Understanding regulation of gene transcription through epigenomics and cistromics : unfolding ones and zeros into (un)folding chromatin

Understanding regulation of gene transcription through epigenomics and cistromics:

Unfolding ones and zeros into (un)folding chromatin

Understanding regulation of gene transcription through epigenomics and cistromics

Unfolding ones and zeros into (un)folding chromatin

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit Maastricht,
op gezag van de Rector Magnificus,
Prof. mr. G.P.M.F. Mols,
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen
op donderdag 21 juni 2012 om 10.00 uur

door

Michiel Emanuel Adriaens
Geboren te Heerlen op 18 December 1983

Promotor:
Prof. Dr. Frederik-Jan van Schooten

Copromotor:
Dr. Chris Evelo

Beoordelingscommissie:
Prof. Dr. Frans Ramaekers (chairman)
Dr. Danyel Jennen
Prof. Dr. John Mathers (University of Newcastle, Newcastle upon Tyne, United Kingdom)
Prof. Dr. Michael Müller (Universiteit van Wageningen, Wageningen, Nederland)
Prof. Dr. Harald Schmidt

The study presented in this thesis was performed within NUTRIM School for Nutrition, Toxicology and Metabolism, which participates in the Graduate School VLAG (Food Technology, Agrobiotechnology, Nutrition and Health Sciences), accredited by the Royal Netherlands Academy of Arts and Sciences.

Printed by: Proefschriftmaken.nl || Uitgeverij BOXPress 's-Hertogenbosch
Published by: Uitgeverij BOXPress, 's-Hertogenbosch

Cover design by Michiel E. Adriaens.

Table of Contents

Chapter 1

General Introduction

To truly understand biology, we need to study it as a system [1]. From a technical perspective, a stable system is characterized by feedback loops and a large degree of redundancy, and biological systems are no exception. Changing a single input, such as the activity of a protein due to a mutation, will generally have only a small impact on the observed phenotype of the biological system, but will affect several biological pathways regardless. This becomes complicated when studying complex traits, such as cancer, which is characterized by a global deregulation of genetic and epigenetic processes, affecting a whole network of intertwined biological pathways [2]. The same goes for nutrition, as food consists of complex mixture of various bioactive compounds, influencing a plethora of processes [3,4]. The only way to unravel such complex interactions, is to use so-called systems biology approaches. Technology wise, this requires high-throughput genome wide methods. No longer focusing on measurements of single entities, biological research now encompasses measurements on multiple molecular levels using so-called omics approaches, to study gene transcription [5], protein and metabolite levels [6], protein-protein interactions [7], protein-DNA interactions [8], genetic variation [9] and many other levels in a single integrative biological framework. With these approaches, not only the amount of data in biological research has exploded, but also the complexity thereof, severely hampering biological interpretation. The major challenge of bioinformatics has been to develop computational approaches to distill biology from this broth of ones and zeros and as such has become one of the cornerstones in modern systems biology research.

Arguably, the most successful omics technology of the last decade has been transcriptomics, to study gene expression in a genome-wide setting [10]. Yet, more and more we realize that we are not only interested in identifying changes in gene expression between conditions, but also in the complex regulatory events behind such changes. This thesis focuses on the bioinformatics challenges of studying regulatory events originating from the cistrome and the epigenome. The cistrome is defined as the complete genome-wide set of cis-acting target sites on the DNA, such as transcription factor binding sites, of a trans-acting factor, such as a transcription factor. The complete genome-wide set of epigenetic marks is known as the epigenome. Such epigenetic marks are heritable chromatin modifications other than changes in the actual underlying DNA sequence, that influence gene expression and more generally phenotype. The study of the cistrome and epigenome in biological systems in a genome-wide fashion using high-throughput technology, is referred to as cistromics and epigenomics, respectively [2, 11, 12].
Epigenetic marks influence gene expression by remodeling the chromatin to be in either an open active state, known as euchromatin, or a closed, densely packed silenced state known as heterochromatin [13]. Epigenetic marks do not only impact the condition of a biological system, but since these marks are intrinsically plastic, they are also affected by impacts on the biological system, such as disease and the environment. It has been suggested that if such impacts are severe, for example during a long period of famine [14], they are passed on to future generations, thereby contributing to the adaptability of

organisms. In this view, changes in epigenetic marks are considered “epigenetic scarring”: significant events in life leave their mark on the genome for future generations.

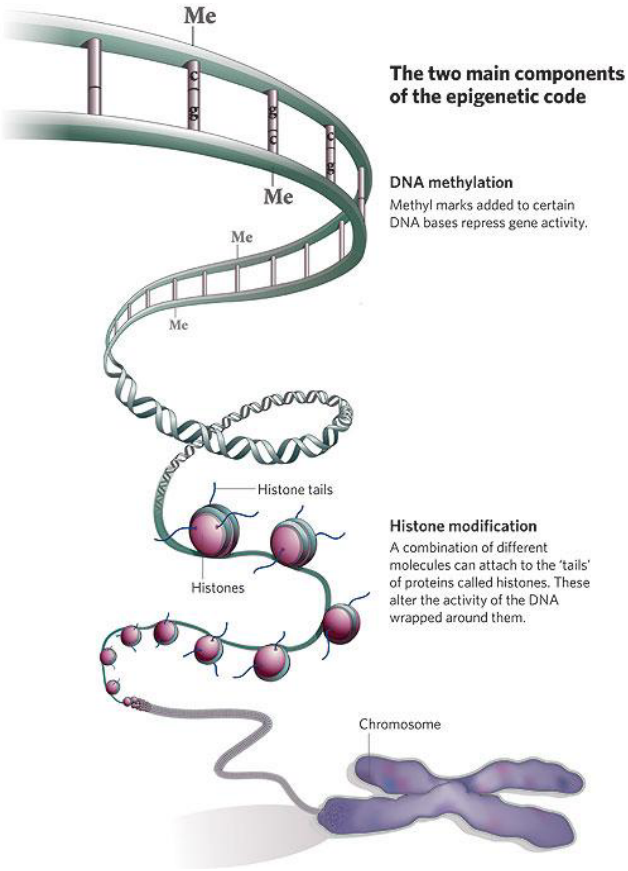


Figure 1: *An overview of Epigenetic Mechanisms (Originally published in Nature [13], used with permission).*

The most studied epigenetic marks are DNA methylation and modification of histone tails (**figure 1**). DNA methylation occurs mostly at CpG di-nucleotides, adding a methyl-group to cytosine to form 5-methylcytosine. Loci with a large amount of CpGs are known as CpG islands, which are overrepresented in regions in the DNA associated to genes. When a CpG island becomes methylated, it causes the nearby gene to become silenced by impeding the binding of required transcription factors and co-regulators. This can occur either directly by decreased binding affinity on methylated CpGs, or by a more permanent mechanism based on attracting methyl-CpG-binding domain proteins (MBD), that subsequently recruit chromatin remodeling complexes, leading to histone modifications that remodel the

chromatin to become silenced. This illustrates that DNA methylation and histone modifications are part of the same larger epigenetic mechanism.
Histones are proteins that act as spools around which DNA winds in segments of 147 base pairs, to form structural units called nucleosomes. A single histone is a complex consisting of histone sub-units known as H1/H5, H2A, H2B, H3, and H4. These sub-units have tails that extend from the complex and are susceptible to chemical modifications (**figure 1**). These histone modifications influence gene transcription, by inducing a specific chromatin state. The canonical modification associated with active chromatin and active transcription is tri-methylation of lysine 4 on histone H3 (H3K4me3). This modification is required at only a few nucleosomes surrounding the transcription start site of a gene to enable transcription. Conversely, tri-methylation of lysine 27 of histone H3 (H3K27me3) is associated with silenced chromatin and silencing of gene transcription, and tends to affect a large number of nucleosomes to cover the entire locus of a gene. This spreading out of H3K27me3 modifications is known as “blanketing”.

The foremost technology to study the epigenome and the cistrome is based on combining chromatin immunoprecipitation with microarrays (ChIP-on-chip [15]) or more recently high-throughput sequencing (ChIP-seq [16,17]). In short, pieces of chromatin bound by a protein or with a specific epigenetic mark, are extracted by using an anti-body (**figure 2 left**). The resulting enriched sample, commonly referred to as the experimental DNA sample, is purified thereby creating a sample consisting of short pieces of DNA. These pieces of DNA are then hybridized to a microarray containing complementary probes for either the whole genome or specific regions of interests, such as gene promoters, or the sample can be analyzed directly using high-throughput sequencing. Using this technology together with bioinformatics tools, genome-wide maps can be constructed of enrichment for specific DNA interacting proteins, histone modifications or DNA methylation, as long as a suitable anti-body is available.
For DNA methylation research, an anti-body is used that directly targets methylated CpGs and the procedure is referred to as methylation dependent immunoprecipitation (MeDIP [18]). Other options for DNA methylation include using methylation specific restriction enzymes, such as in the McrBC approach [19] (**figure 2 right**), or more recently, by using an antibody against MBD protein bound to methylated DNA regions [20]. Each approach has its own strengths and weaknesses. MeDIP for example is biased towards highly methylated, CpG rich regions, such as CpG islands and less sensitive in picking up high methylation in CpG poor regions, while the MBD approach is more sensitive to highly methylated regions with moderate CpG density [21]. As such, these technologies are complementary and can be used conjointly in an experiment [22].

What is clearly lacking in the field are standardized data processing and interpretation procedures for epigenomics and cistromics data and the lack of a comprehensive toolset to perform these analyses in. Standardization is absolutely essential to support the dissemination of this technology outside the confined space of the academic laboratory and pave the way towards clinical applications [23]. Hence,

the main aim of this thesis is to improve cistromics and epigenomics data interpretation through development of standardized analysis approaches, hereby addressing existing issues in every step of the analysis process, from raw data preprocessing to biological interpretation of the results. In each chapter, a specific issue is addressed and solutions proposed.

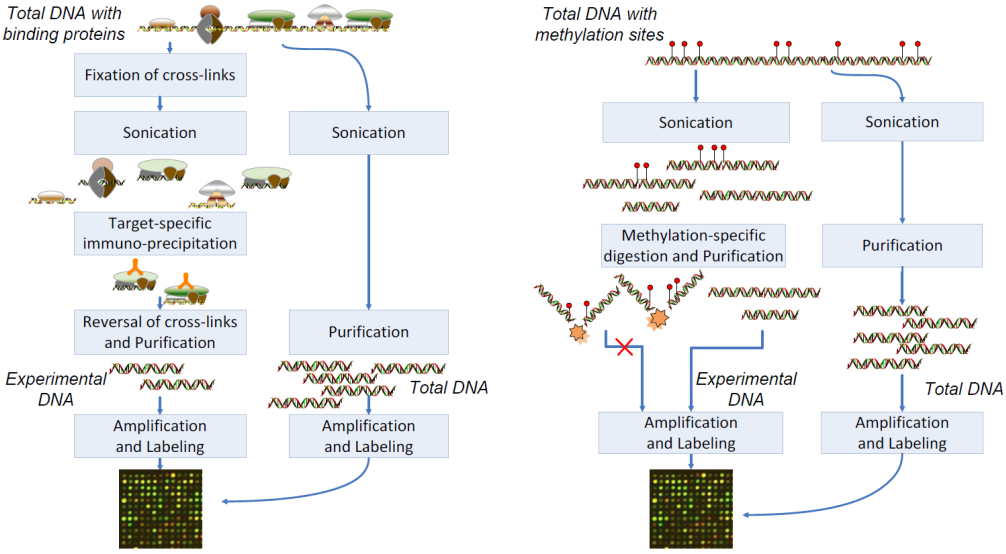


Figure 2: *An overview of two protocols to perform cistromics and epigenomics analyses. The general assay for ChIP and MeDIP studies is shown on the left, whereas the assay for DNA methylation based on methylation specific restriction enzymes is shown on the right.*

Chapter 2 formulates the requirements of a cistromics and epigenomics work-flow and the novel enrichR toolkit that enables to perform such analyses swiftly and thoroughly. As with any omics technology, cistromics and epigenomics data are inherently noisy. Due to this nature, microarray based approaches are intended for exploratory data analysis to generate hypotheses to validate in the lab. To enhance interpretation, ideally multiple lines of evidence are used, combining measurements from various omics technologies as well as multiple in silico analysis approaches. For cistromics and epigenomics we are interested in regulation of gene transcription. Thus, the primary requirement for a meaningful data interpretation is integration with transcriptomics data. Additionally, we are interested in uncovering cis-acting motifs present in the underlying DNA sequences of the identified enriched regions that potentially direct the binding of DNA interacting proteins. The software of choice for any omics data analysis is Bioconductor [24], which is a comprehensive, open-source collection of bioinformatics analysis packages. EnrichR has been built using this framework and in accordance with its core philosophy is available as open-source.

The first step in any analysis using high-throughput technology is pre-processing of the raw measurements to correct for variance that is of technical rather than biological nature. In chapter 3, preprocessing issues for ChIP-on-chip and MeDIP-on-chip microarray data are discussed. Although derived from transcriptomics technology, there are key differences in the analysis of ChIP-on-chip and MeDIP-on-chip microarray data. Most commonly performed on two-channel microarray technology, one channel contains the enriched sample, while the other contains a total DNA sample that contains all the sequence content from the studied biological system as a reference. The ratio between the channels is used as the measure to determine enriched regions in the genome, corresponding to DNA interacting protein binding sites or sites with specific epigenetic modifications. This setup is different from two-channel transcriptomics microarrays, where both channels contain amplified transcript samples usually corresponding to two different conditions. The general assumption in transcriptomics studies is that the majority of genes are unchanged between conditions [25,26,27], and hence the majority of spots on a microarray will have a comparable signal between channels as well. In epigenomics and cistromics studies however, this assumption does not hold since the samples comprising the two channels are so fundamentally different, which suggests that many approaches developed for two-channel transcriptomics microarray data may not be suitable. To determine whether this is indeed the case, we evaluated the performance of several well-known transcriptomics normalization strategies when used for ChIP-on-chip and MeDIP-on-chip technology. The results are discussed in chapter 3, where we see that T-quantile normalization applied separately on the channels and Tukey's biweight scaling are the only approaches consistently yielding the best results, where in contrast, popular normalization approaches, like quantile and LOWESS normalization, impact the reliability of the downstream analysis substantially.

In chapter 4 we apply the lessons learned from ChIP-on-chip and MeDIP-on-chip technology to construct an optimal processing protocol for H3K27me3 histone modification ChIP-seq data. There are specific challenges for studying this histone modification in dynamical biological systems, in which a considerable amount of epigenetic changes are expected to occur, both in location as well as in total amount. Firstly, enrichment finding algorithms for ChIP-seq data are optimized for locating sharply defined peaks [28], but due to blanketing [29,30,31], the H3K27me3 enrichment signal is spread out instead. Secondly, like ChIP-on-chip, MeDIP-on-chip and any other high-throughput technology, ChIP-seq data requires normalization to correct for technical bias and enable quantitative comparisons between samples. When studying dynamical biological systems, the only suitable approach is scaling the data based on regions in the genome where enrichment is stable between conditions. It is however difficult to define a priori which regions are prone to have such stable enrichment, as this depends heavily on the biological system studied [32]. With these difficulties in mind, we have developed a standardized protocol for the processing of H3K27me3 ChIP-seq data. The protocol enables robust detection of H3K27me3 blanketing and allows for quantitative data comparison. As such, our protocol complements previous efforts to create the first fully standardized analysis pipeline for H3K27me3 enriched ChIP-seq data. We have used this protocol to map enrichment of H3K27me3 histone modifications genome-wide in breast cancer cells that are

exposed to hypoxic conditions, serving as a model for dynamic effects occurring in response to hypoxia in solid tumors.

After raw data has been processed, biology finally comes into view. In chapter 5 we compare the entries for several pathways associated with fatty acid metabolism and assess their quality by expanding them with current literature and subsequently having them curated by experts. There is a tremendous body of biological knowledge already available in biological databases in the form of relations between bioactive molecules. This information can be stored as single biochemical interactions [33] or as networks of biochemical interactions [34, 35]. When such a network comprises a more or less defined biological process with a sense of direction [36], it is called a pathway. An important step in data interpretation is integrating knowledge from biological databases. One of the most common approaches for this is pathway analysis, which maps gene related data, such as expression data, onto existing biological pathways, allowing a straight-forward visual interpretation [37]. Pathways such as the one in **figure 3** are static representations of biological processes. However, the quality of the content depends highly on the knowledge of its creators and the time since it was last updated. During the assessment presented in chapter 5, we discovered major differences between the content of databases, and created updated, comprehensive pathways of fatty acid metabolism which can be found on WikiPathways [34].

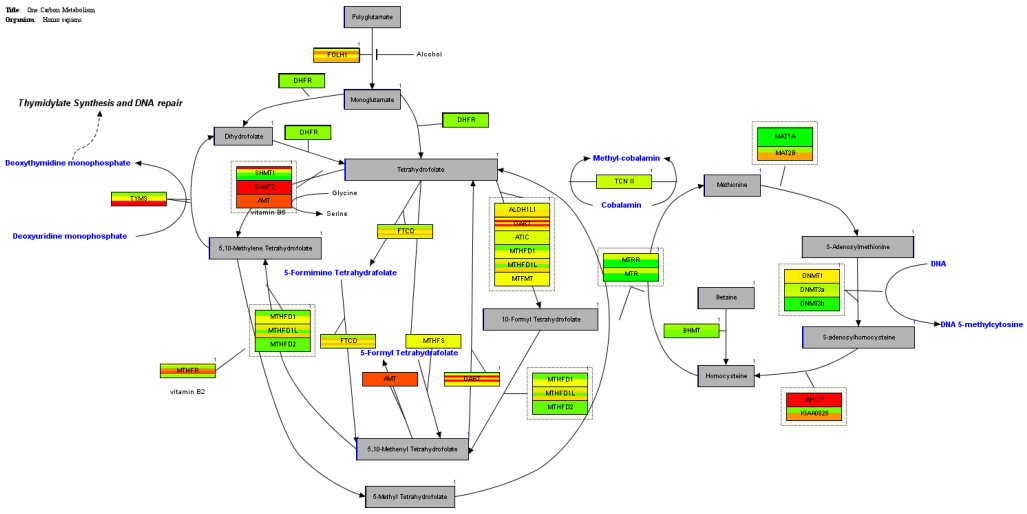


Figure 3: *An example of a pathway with mapped transcriptomics data. Red colored gene boxes indicate up-regulation of the associated transcript, while green indicates down regulation. The pathway was downloaded from WikiPathways (http://wikipathways.org/index.php/Pathway:WP241) and data was mapped using Pathvisio [37].*

Chapters 6 and 7 serve as demonstrations in a biological setting of the approaches discussed in chapters 2, 3 and 4. Chapter 6 describes the analysis and results of an estrogen-receptor α (ER-α) ChIP-on-chip dataset aimed at the dissection of the molecular mechanisms of the mitogenic action of estrogen in the human endometrium and the breast. Chapter 7 meanwhile describes the results of a ChIP-seq data analysis on the enrichment of activating H3K4me3 and silencing H3K27me3 histone modifications. The studied system is an MCF7 breast cancer cell line that is exposed to hypoxic conditions, as a model of effects occurring in the cores of solid tumors. Cancer cells are characterized by many epigenetic deregulations. Understanding the changes in histone modifications in such a dynamical system is key to understanding the molecular mechanisms underlying cancer and for the development of future cancer treatment. These applications show the importance of the developed approaches and the power of integrating cistromics and epigenomics technology in systems biology research.

References

1. Lazebnik I: [Can a biologist fix a radio, or what I learned while studying apoptosis]. Advances in gerontology 2003, 12:166-171.
2. Esteller M: Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet 2007, 8(4):286-298.
3. de Roos B: Proteomic analysis of human plasma and blood cells in nutritional studies: development of biomarkers to aid disease prevention. Expert review of proteomics 2008, 5(6):819-826.
4. de Roos B, McArdle HJ: Proteomics as a tool for the modelling of biological processes and biomarker development in nutrition research. The British journal of nutrition 2008, 99 Suppl 3:S66-71.
5. Martin JA, Wang Z: Next-generation transcriptome assembly. Nature reviews Genetics 2011, 12(10):671-682.
6. Patterson SD, Aebersold RH: Proteomics: the first decade and beyond. Nature genetics 2003, 33 Suppl:311-323.
7. Bonetta L: Protein-protein interactions: Interactome under construction. Nature 2010, 468(7325):851-854.
8. Gilchrist DA, Fargo DC, Adelman K: Using ChIP-chip and ChIP-seq to study the regulation of gene expression: genome-wide localization studies reveal widespread regulation of transcription elongation. Methods (San Diego, Calif) 2009, 48(4):398-408.
9. Cotton RGH, Auerbach AD, Axton M, Barash CI, Berkovic SF, Brookes AJ, Burn J, Cutting G, den Dunnen JT, Flicek P et al: GENETICS. The Human Variome Project. Science (New York, NY) 2008, 322(5903):861-862.
10. Schena M, Heller RA, Theriault TP, Konrad K, Lachenmeier E, Davis RW: Microarrays: biotechnology's discovery platform for functional genomics. Trends in biotechnology 1998, 16(7):301-306.

11. Lupien M, Brown M: Cistromics of hormone-dependent cancer. Endocrine-related cancer 2009, 16(2):381-389.
12. Flintoft L: Epigenomics: Reprogramming in transition. Nature reviews Genetics 2011, 12(8):522.
13. Qiu J: Epigenetics: unfinished symphony. Nature 2006, 441(7090):143-145.
14. Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES, Slagboom PE, Lumey LH: Persistent epigenetic differences associated with prenatal exposure to famine in humans. Proceedings of the National Academy of Sciences of the United States of America 2008, 105(44):17046-17049.
15. Zheng M, Barrera LO, Ren B, Wu YN: ChIP-chip: Data, Model, and Analysis. Biometrics 2007, 63(3):787-796.
16. Park PJ: ChIP-seq: advantages and challenges of a maturing technology. Nature reviews Genetics 2009, 10(10):669-680.
17. Mardis ER: ChIP-seq: welcome to the new frontier. Nature methods 2007, 4(8):613-614.
18. Mohn F, Weber M, Schübeler D, Roloff T-C: Methylated DNA Immunoprecipitation (MeDIP). Methods Mol Biol 2009, 507:55-64.
19. Ordway JM, Bedell JA, Citek RW, Nunberg A, Garrido A, Kendall R, Stevens JR, Cao D, Doerge RW, Korshunova Y et al: Comprehensive DNA methylation profiling in a human cancer genome identifies novel epigenetic targets. Carcinogenesis 2006, 27(12):2409-2423.
20. Serre D, Lee BH, Ting AH: MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. Nucleic acids research 2010, 38(2):391-399.
21. Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, Downey SL, Johnson BE, Fouse SD, Delaney A, Zhao Y et al: Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nature biotechnology 2010, 28(10):1097-1105.
22. Li N, Ye M, Li Y, Yan Z, Butcher LM, Sun J, Han X, Chen Q, Zhang X, Wang J: Whole genome DNA methylation analysis based on high throughput sequencing technology. Methods (San Diego, Calif) 2010, 52(3):203-212.
23. MAQC-II: analyze that! Nat Biotech 2010, 28(8):761-761.
24. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J et al: Bioconductor: open software development for computational biology and bioinformatics. Genome biology 2004, 5(10):R80.
25. Yang YH, Dudoit S, Luu P: Normalization for cDNA microarray data. Optical Technologies and Informatics 2001, 4266:141-152.
26. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielser HBr, Saxild H-H, Nielsen C, Brunak Sr, Knudsen S: A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome biology 2002, 3(9):research0048.

27. Dudoit S, Yang YH, Callow MJ: Statistical methods for identifying differentially expressed genes in replicated cDNA microarray. Statistica Sinica 2002, 12:111-139.
28. Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones SJM: FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics (Oxford, England) 2008, 24(15):1729-1730.
29. Bracken AP, Dietrich N, Pasini D, Hansen KH, Helin K: Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. Genes & development 2006, 20(9):1123-1136.
30. Pietersen AM, van Lohuizen M: Stem cell regulation by polycomb repressors: postponing commitment. Current opinion in cell biology 2008, 20(2):201-207.
31. Young MD, Willson TA, Wakefield MJ, Trounson E, Hilton DJ, Blewitt ME, Oshlack A, Majewski IJ: ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. Nucleic acids research 2011, 39(17):7415-7427.
32. Huang W, Umbach DM, Vincent Jordan N, Abell AN, Johnson GL, Li L: Efficiently identifying genome-wide changes with next-generation sequencing data. Nucleic acids research 2011, 39(19):e130.
33. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B et al: Reactome: a database of reactions, pathways and biological processes. Nucleic acids research 2011, 39(Database issue):D691-697.
34. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C: WikiPathways: pathway editing for the people. PLoS biology 2008, 6(7):e184.
35. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research 2011.
36. Cary MP, Bader GD, Sander C: Pathway information for systems biology. FEBS letters 2005, 579(8):1815-1820.
37. van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S, Conklin BR, Evelo C: Presenting and exploring biological pathways with PathVisio. BMC Bioinformatics 2008, 9(1):399.

Chapter 2

Systems biology approaches for ChIP-on-chip and DNA methylation microarray data

Michiel E. Adriaens[1], Wibowo Arindrarto[1], Andrea Romano[2,3], Lars M.T. Eijssen[1] and Chris T.A. Evelo[1]

[1] *Department of Bioinformatics – BiGCaT, Maastricht University, Maastricht, The Netherlands*
[2] *GROW, School for Oncology and Developmental Biology, Maastricht University, Maastricht, The Netherlands*
[3] *Department of Obstetrics and Gynaecology, Maastricht University, Maastricht, The Netherlands*

Keywords: *ChIP-on-chip, MeDIP-on-chip, bioinformatics, systems biology, Bioconductor.*

Publication: Manuscript in preparation.

Abstract

Advances in epigenomics and cistromics have made it possible to study gene transcription regulation on a genome wide level. ChIP-on-chip is used to determine binding locations of DNA interacting proteins genome wide, be it transcription factors or histones with specific modifications. The same microarray technology and analogous protocols can be applied to detect the DNA methylation status of thousands of CpG rich regions in one go. The DNA enrichment data arising from these technologies opens up tremendous opportunities for unraveling gene transcription regulation. But no matter how stringent the statistics, validation of potential regulated genes is essential, using multiple lines of evidence from various bioinformatics analysis approaches and additional assays to robustly characterize regulated targets. Since such a work flow requires a flexible and extendible framework, we used R and the Bioconductor framework to develop enrichR, a toolbox for integrative analysis of ChIP-on-chip and DNA methylation microarray data, optimized for the popular NimbleGen platform. The software is fully open-source with a focus on high-level functions to accommodate users from any experience level, enabling a full biological interpretation in minutes instead of days. As an illustration of enrichR, we have performed an integrative analysis of a public estrogen-receptor α (ER-α) ChIP-on-chip dataset and corresponding transcriptomics dataset, aimed at dissection of the molecular mechanisms of the mitogenic action of estrogen in the human endometrium and the breast.

Introduction

Over the years, it has become clear that to understand biology, we need to study it as a system [1]. Many genome wide technologies have been developed to study biological systems, of which microarray technology is the most established. Although microarray technology to measure gene transcription has been around since the turn of the millennium, to characterize a system, we need to characterize its regulators. Chromatin immunoprecipitation followed by array hybridization (ChIP-on-chip) [2] or, more recently, followed by high-throughput sequencing (ChIP-seq), is used to determine genome wide binding locations of DNA interacting proteins, such as transcription factors or histones. The same microarray technology and analogous protocols can be applied to detect the DNA methylation status of thousands of CpG rich regions, such as CpG islands, in one go [3]. The data arising from these technologies, henceforth referred to as DNA enrichment data, open up tremendous opportunities for unraveling gene transcription regulation and there are many studies that, using such approaches, have given relevant contributions to our understanding of the molecular mechanisms underlying gene transcription, cell differentiation and cell lineage maintenance [4,5,6,7,8].

In spite of this popularity, analysis approaches for DNA enrichment microarrays are not as standardized as for instance for transcriptomics microarray analysis. There are several reasons for this. First and foremost, the underlying technology and study designs of these types of microarray analyses are extremely diverse. The second reason is that most popular technology for DNA enrichment analysis is

based on two-channel microarray technology. This means that there will always be two samples on each array, with the golden standard being input DNA in one channel, comprising all DNA sequences present in a sample, and an immunoprecipitated DNA sample in the other, which is enriched for regions of interest, such as binding locations of a DNA interacting protein. Depending on the research question, other designs may yield more robust results for less costs. If for instance the goal is to identify binding regions that differ between two samples, or differences in degrees of methylation, hybridizing them together on one chip but in different channels is a good approach [9].

Apart from these technical issues, there are specific problems in the analysis of DNA enrichment microarray data. With respect to data normalization, since samples on this two-color technology differ to such a large extent, standard microarray normalization procedures can be very destructive on the power to identify enrichment and differences therein [10]. Although this is less of an issue for the identification of strongly enriched regions, this becomes a major limitation when small differences or moderate enrichment is expected, which is frequently the case for transcription factors which only mildly and transiently associate to chromatin. Destructive normalization will hamper the identification of such differences.

This problem is intensified by the fact that the number of expected enriched regions is in most cases unknown and may differ to considerable extents between experimental conditions. Therefore, a suitable positive or reference control is missing. Most approaches to identify enriched regions are within-sample approaches, leading to relative enrichment scores, which are difficult to interpret [11,12]. Since there is a definite but unknown correlation between FDR values and relative amount of binding or DNA methylation, depending on the amount of enriched regions present in the sample, this paves the way for arbitrary cut-offs to assess significant enrichment.

With respect to the last step in the analysis process, biological interpretation, enrichment and differences therein may not have a measurable effect on gene transcription, even when this difference is classified as statistically significant. In general, DNA methylation is known to silence gene transcription, which would imply a negative correlation between changes in DNA methylation and changes in gene transcription. This model is too simple though, as a decrease in DNA methylation of a gene's promoter does not impose that gene to become activated, as a suitable set of transcription factors is still needed, as is an active chromatin structure, guided by the necessary histone modifications. On top of that, small changes in efficiency of protein binding or DNA methylation across a tissue can be completely counteracted by changes in the levels of available transcription co-regulators. If for instance a gene is hypermethylated in a specific tissue, this means that the average degree of methylation of this gene across multiple copies of that gene is increased. This implies that the gene is not silenced in all cells of such a sample and are exposed for regulation by transcription factors in cells with unmethylated DNA.

DNA interacting proteins can repress gene transcription, activate it or both, depending on available co-regulators that are assembled in the multi-protein complexes associated with chromatin [13]. Since inactive chromatin hampers the formation of DNA protein complexes, histone modifications are usually

less of an issue in the process of target identification in ChIP-on-chip experiments. This changes dramatically of course, when using ChIP technology to measure such histone modifications. Then similar effects occur as with DNA methylation microarray data, where the effect on gene transcription of a change in a particular activating or silencing modification is subject to availability of required co-regulators.

These observations imply that results DNA enrichment experiments on itself are extremely ambiguous and interpretation of such results is severely hampered by lack of additional lines of evidence. Hence, more than with any other type of technology, systems biology approaches are essential to get the biology out of DNA enrichment data. Therefore, a work flow is needed that implements multiple lines of evidence from various bioinformatics analysis approaches and additional assays to robustly characterize regulated targets. We have designed such a work flow to maximize biological output of DNA enrichment data, and to do it fast and concise (**figure 1**).

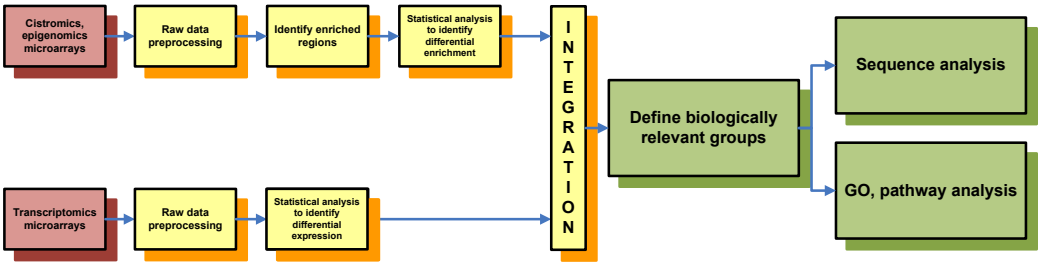


Figure 1: *Integrative work flow for DNA enrichment microarray data: from raw data (red), through data analysis (yellow), to biological interpretation (green).*

The essential step is integration with transcriptomics data, ideally running a gene expression assay in parallel to the DNA enrichment analysis. Alternatively, online data repositories such as ArrayExpress [14] house over a decade of transcriptomics datasets, which therefore bound to include a suitable dataset for integration. Based on the integration, genes are split in multiple groups, depending on their relative expression and DNA enrichment trends. This step is the least generic in the entire work flow, since it depends on the biology behind the experiment and is therefore strongly data driven. An example is given in **figure 2** where the differences in DNA methylation between group A and group B are plotted against the differences in gene expression of all significantly regulated genes, leading to four distinct groups. These groups can be biologically characterized by performing biological process enrichment analysis, either using Gene Ontology enrichment analysis [15] or using pathway analysis tools [16], and based on the results sub-divided into groups. Transcription factors need co-regulators to form complexes and influence gene transcription and hence the binding of a single transcription factor is an uninformative

indicator of regulation [8]. Hence, a subsequent essential in silico validation step when studying transcription factor binding is sequence motif analysis of enriched regions [17,18,19], to identify discriminating motifs between groups, such as the overrepresentation of a set of transcription factor binding sites in one group compared to another. For DNA methylation studies, such motif analysis would seem less informative, but specific classes of CpG rich regions with tissue and phenotype specific methylation are often characterized by enrichment for specific motifs, suggesting that cis-regulatory elements may facilitate differential methylation at such CpG islands [20,21,22].

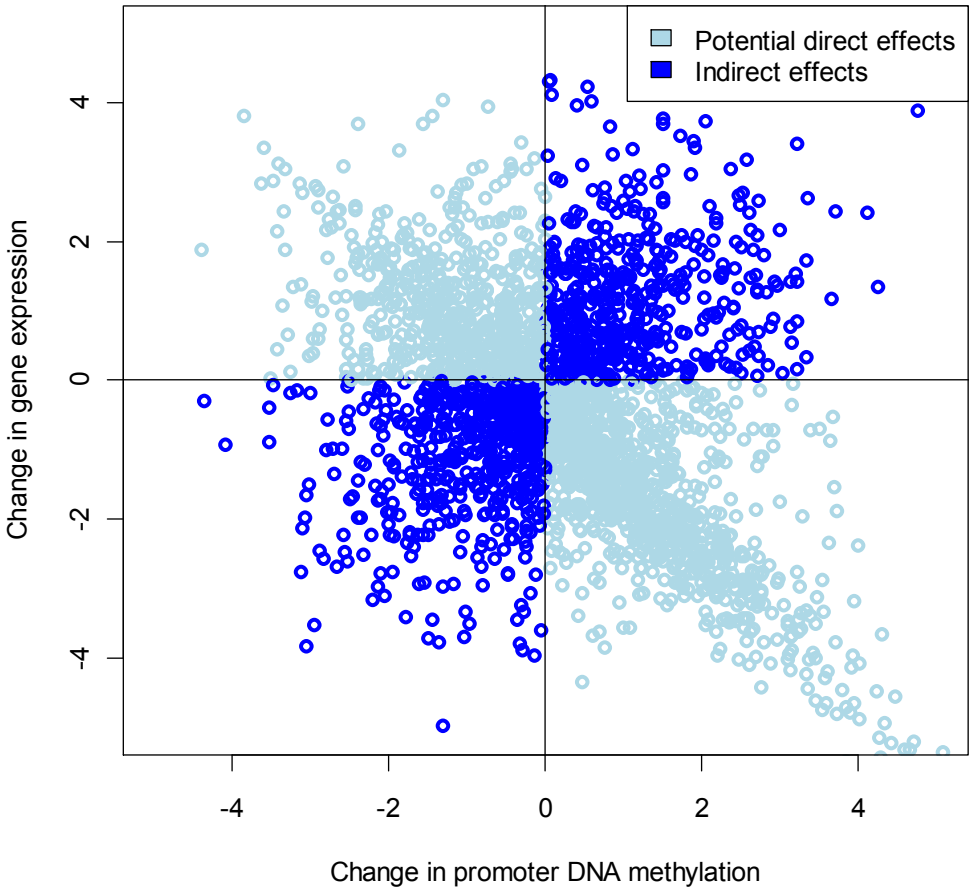


Figure 2: *Scatterplot of changes in DNA methylation versus changes in gene expression of simulated DNA enrichment data.*

The work flow of **figure 1** imposes several requirements on the software used to perform the analyses. It should be flexible and well suited to multiple types of data from multiple technologies. Also, it needs to be able to retrieve sequence information, since probes are short, often non-overlapping, so gaps need to be

filled to be able to perform sequence motif analysis. A tool that lends itself well to integrative systems biology approaches is R [34]. R is an open source statistical analysis environment and programming language. It is easily extended with additional functionality through packages. The Bioconductor project is a great example of this, which is a collection of packages to suit all bioinformatics and biostatistics needs. R combined with Bioconductor is the de facto open source standard for genome wide data analysis, which makes R a logical choice for ChIP and DNA methylation microarray analysis. There are however many minor and major annoyances in R. Although it is flexible, it is also unstable and ever changing. Packages that your analysis work flow relied on previously may be altered beyond recognition in an updated version. In addition, some essential functionality for ChIP and DNA methylation microarray analysis is often missing or suffering from cluttered interfaces, such as retrieving sequences of enriched regions from a sequence database. It is like an analog synthesizer: it can make any sound on the planet, but enabling it to do so requires an advanced level of programming knowledge.
With this in mind, we have created enrichR, a toolbox to facilitate any type of enrichment analysis, be it ChIP-on-chip, DNA methylation or any other type of enrichment microarray technology. It combines several well established Bioconductor packages and adds to them what we found missing, creating a framework for robust multiple validation analysis. The package wraps around a collection of well established and stable Bioconductor packages. To accommodate users from any experience level, we optimized and automated our toolkit for the popular NimbleGen platform, creating several high-level functions. Most importantly, we focused on promoter report data, the data received back from NimbleGen, enabling to perform a full biological interpretation in minutes instead of days. Our toolkit has formed the basis for the bioinformatics analysis of several papers [9,13,23] and has been used extensively for education purposes [24].

To demonstrate our work flow and toolkit, we reanalyzed a study on estrogen receptor α targets, aimed at dissection of the molecular mechanisms of the mitogenic action of estrogen in the human endometrium and the breast [13]. Estrogens are one of the most frequently prescribed drugs worldwide [25]. All major actions of estrogen are mediated by the estrogen receptors of which ER-α is responsible for proliferation and cell homeoastasis in gynaecologic tissues. Ligand activated ER-α binds to promoters of target genes and depending on the availability of co-activators or co-repressors, the transcription of the target genes is induced or repressed [13, 26, 27].
Compounds with estrogen agonistic action protect against osteoporosis, are beneficial for a number of physiological functions (cardiovascular, lipid profile) and are used in hormone replacement therapy (HRT). On the contrary, estrogen antagonists are used in breast cancer therapy. These drugs, however, lead to unwanted side effects, because agonists used in HRT are mitogenic factors in the breast and the endometrium and may cause endometrial or breast cancer, whereas antagonists increase the risk for osteoporosis and cardiovascular events. The pharmaceutical industry has developed compounds known as selective estrogen receptor modulators (SERMs [28]), which, in contrast to pure agonists or

antagonists, display tissue-specific agonistic/antagonistic activities. Therefore, potentially, the ideal SERM would be antagonists in gynaecologic tissues and the breast, and agonist in the bone and the cardiovascular system. However, such a SERM is not available yet, and the use of these medications is still confronted with important clinical problems. For instance, tamoxifen, which is used to treat breast cancer with excellent results, induces cell proliferation in the uterus, thus increasing the risk for benign and malignant endometrial lesions. The inability to predict the tissue specific response to estrogens and SERMs is due to the fact that the mechanism of action of ER-α and SERMs is poorly understood. Therefore, the recruitment of distinct co-regulators by tamoxifen-activated ER-α in the breast compared to the endometrium would lead to distinct transcriptional activation of target genes, hence distinct agonistic/antagonistic responses. Therefore, by dissecting differential co-regulator recruitment, it will be possible to predict the action of novel SERMs in the human endometrium and other tissues or to identify patients who will benefit from hormonal therapies.

Results

As an illustration of enrichR and our work flow, we have performed an integrative analysis of an estrogen-receptor α (ER-α) ChIP-on-chip dataset aimed at dissection of the molecular mechanisms of the mitogenic action of estrogen in the human endometrium and the breast [13]. We have combined the original ChIP-on-chip data with public transcriptomics datasets, and defined distinct groups of targets based on the integrated results. We analyzed these groups separately, providing them with a biological context by performing biological process enrichment analysis using Gene Ontology and exploring the presence of discriminative sequence motifs in and surrounding the enriched regions and promoters of target genes.

The study consisted of a ChIP-on-chip assay. Briefly, estrogen responsive T47D breast cancer cells were treated with 17-β-estradiol or with vehicle alone (ethanol). Chromatin-immunoprecipitation using an ER-α specific antibody was performed 50 minutes after induction, when gene promoters are targeted by the activated receptor. Precipitated DNA sequences were subsequently amplified, labeled and hybridized to a NimbleGen Human HGS17 minimal promoter array containing 24134 putative gene promoters. Targets were identified as previously described [13].

Next, we integrated the results with gene expression data from a study investigating gene expression in T47D cells in response to 17-β-estradiol stimulation [29]. Briefly, T47D cells were cultured in culture medium which was supplemented with 17-β-estradiol. Gene expression profiles were taken at 0, 1, 2, 4, 8, 12 and 24 hours, using Affymetrix Human Genome U133A 2.0 GeneChips.

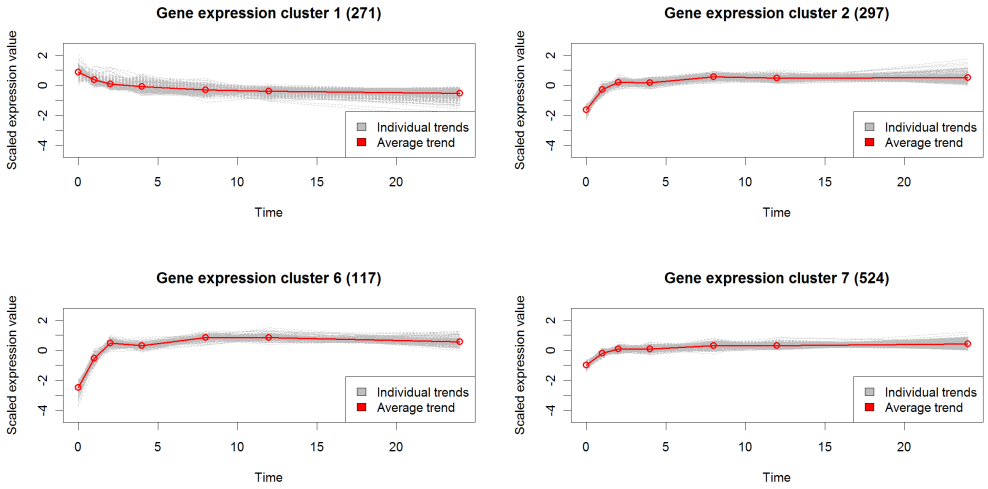


Figure 3: *Clusters of significantly regulated genes, based solely on the transcriptomics data.*

From the transcriptomics results, we created profiles or clusters of significantly differentially regulated genes. From these clusters, we defined two categories that showed the largest difference between time point 0 and 1: one of down-regulated genes (**figure 3, top left**), one of up-regulated genes (**figure 3, others**) and performed Gene Ontology enrichment analysis on this set of genes. Although these clusters will encompass both direct effects by binding of 17-β-estradiol activated estrogen receptor α and indirect effects by downstream regulation, the effects are in any case due to estrogen stimulation, and hence the results from this biological process enrichment are expected to be insightful.

The results show that up-regulated genes are involved mainly in cell cycle and gene transcription regulation, indicating that 17-β-estradiol stimulation indeed leads to stimulation of cell cycle inducers (**table 1**). The down-regulated genes are involved in regulation of metabolism and negative regulation of transcription, indicating that 17-β-estradiol stimulation leads to down regulation of genes involved in metabolism and transcriptional repression (**table 1**). Interestingly, some down-regulated genes are also involved in positive regulation of transcription, indicating silencing of some transcriptional inducers.

To analyze direct ER-α targets more closely, we assessed the expression of targets identified in the ChIP-on-chip study and chose those that show significant changes in expression over time and show a absolute fold change larger than 1.4 from time point 0 hours to time point 1 hour. Using these settings, we identified 47 up-regulated targets and 16 down-regulated targets.

Based on that selection, we retrieved the promoter regions of these genes close to the TSS (500 bp upstream, 100 downstream of TSS) and scanned them for presence of CpG islands and ER-α motifs using cosmo [17]. Using strict settings (CpG ratio of 0.65, GC-content of 55% and a minimum length of

500 bp [30]), we found that of the 47 promoters of up-regulated promoters, 29 contain a CpG island (62%). Of the 16 promoters of the down-regulated targets, 9 contain a CpG island (56%).

	GO term	# genes with GO term	p-value
Up-regulated genes	GO:0006350| transcription	125	<0.01
	GO:0045449| regulation of transcription	117	0.04
	GO:0007049| cell cycle	74	<0.01
	GO:0015031| protein transport	59	<0.01
	GO:0008283| cell proliferation	58	<0.01
	GO:0042981| regulation of apoptosis	53	<0.01
	GO:0006461| protein complex assembly	49	0.01
	GO:0050790| regulation of catalytic activity	48	0.01
	GO:0006412| translation	40	<0.01
Down-regulated genes	GO:0019222| regulation of metabolic process	54	0.03
	GO:0045893| positive regulation of transcription, DNA-dependent	15	0.01
	GO:0045892| negative regulation of transcription, DNA-dependent	13	0.01
	GO:0000122| negative regulation of transcription from RNA polymerase II promoter	9	<0.01
	GO:0007264| small GTPase mediated signal transduction	9	0.03
	GO:0045944| positive regulation of transcription from RNA polymerase II promoter	9	0.01

Table 1: *Gene ontology enrichment results for the clusters of significantly regulated genes, based solely on the transcriptomics data. Results only reported back for categories with more than 5% of the total number of genes.*

ER-α can bind to specific estrogen response elements (EREs) or to other motifs when forming a complex with other transcription factors. The ERE consensus sequence, an inverted repeat of the sequence (GGTCA) separated by 3 bp, rarely occurs in nature [6, 27]; however, the imperfect ERE (GGTCANNNTNNCY) and ERE half-site (AGGTCA) are widely accepted as alternative binding sites [27]. When ER-α forms a complex with other transcription factors, the resulting binding motif depends on which proteins interact directly with the chromatin. Using cosmo, we scanned for motifs containing at least the ERE half-site (AGGTCA), using settings that favor a low-stringency palindromic motif. Interestingly, the results show that up-regulated genes are characterized by a ERE half-site, while the down-regulated genes are characterized by a motif more resembling the ERE palindrome (**figure 4**).

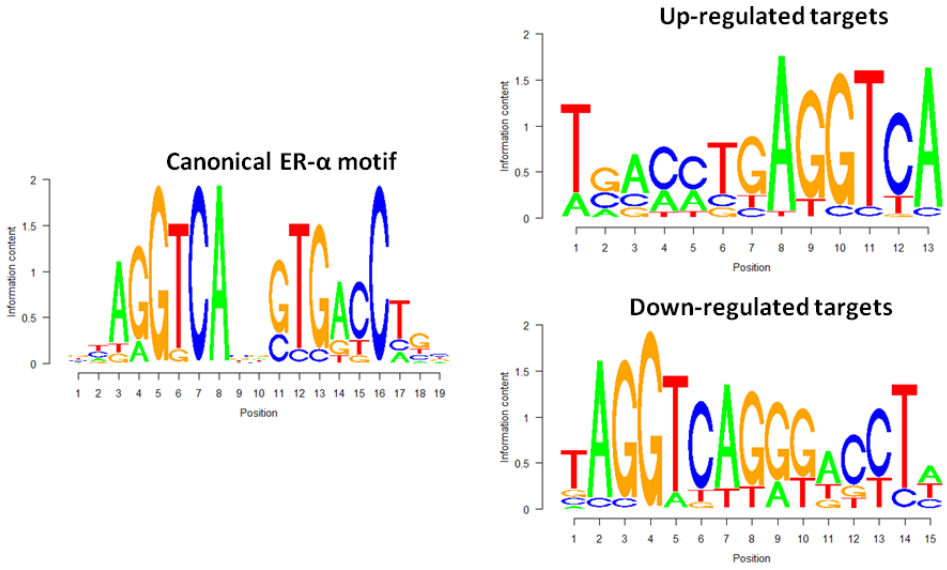


Figure 4: *ER-α motifs: canonical palindromic motif (left) and the overrepresented motifs found in up-regulated and down-regulated targets (right)*

Discussion

The enrichR toolkit presented here consists of a set of tools to ease ChIP-on-chip and DNA methylation microarray analysis and is able to perform complex tasks in just a few lines of code. Galaxy [33] has put emphasis on smaller, easy to integrate tools that together comprise a work flow that is adaptable to any biological question. The enrichR toolkit fits into this trend very well and since Galaxy is fully compatible with R, enrichR can be easily implemented into new and existing Galaxy work flows. Although the enrichR toolkit is designed for epigenomics and cistromics microarray data, the functionality is easily translated to epigenomics and cistromics analyses performed on high-throughput sequencing technology.

Although there are many tools available for DNA enrichment analysis, most of these are directed towards ChIP-on-chip experiments. Examples of these are the Bioconductor [34] packages Ringo [35], which is focused on ChIP-on-chip data analysis for the NimbleGen platform, and Starr [36], which is focused on single-channel Affymetrix ChIP-on-chip data. Examples of stand-alone tools are CisGenome [37], which is suited for both ChIP-on-chip as well as ChIP-seq data, and CoCAS [38], which is again focused on only ChIP-on-chip data. There are also numerous packages for DNA methylation analysis. Examples are the MEDME package [39] in Bioconductor which can estimate DNA methylation percentages for all regions present on an array, but to calculate these, imposes strict requirements on the wet-lab side of the experiment, requiring a fully methylated sample for comparison. BATMAN [40] promises similar things, but sports a Spartan command line interface, requiring extensive knowledge of both Java and database setup to be able to analyze DNA methylation microarray data. Regardless, all these tools are very

powerful on their own, but they all stop when it gets interesting: when we want to go back to biology. Also, the stand-alone tools are hampered by their inability to integrate other types of data or perform additional analyses once enriched regions have been identified.

One of the most important analyses for DNA enrichment data is conserved motif analysis. Such an analysis can be directed or undirected. Directed sequence motif analysis would try to validate identified targets as a having regulatory motifs associated with the transcription factor of interest or some other combination of motifs, while undirected would encompass searching for motifs of interest in all measured regulatory regions and comparing this to the outcome of the DNA enrichment experiment, potentially looking at additional regions that did not make the cut before. This leads to the important observation that instead of filtering out all regions that do not exhibit significant enrichment, all regions should always be reported back. Even regions of low methylation for instance are of interest, and although they are difficult if not impossible to identify by standard enrichment finding procedures, combining such procedures with additional validation steps will enable the identification of such regions of interest and distinguishing them from false positives or background noise. This brings us to the important conclusion that a data driven, multiple validation approach negates the potential for overfitting in model based enrichment finding algorithms [1,12,39,40,31]. The MPEAK algorithm for instance searches for regions that show triangular shaped enrichment, assuming a Poisson distribution around potential transcription factor binding sites [1]. Although this is a sensible approach, putting such a shape constraint on the data can be dangerous, as it does not take probe sequence or spacing effects or other technical or biological effects into account. Such a biological effect might for instance be that the transcription factor of interest binds in a complex of other transcription factors to the DNA, increasing the total length of the binding site and truncating or skewing the triangular shape. This is for instance the case when studying certain histone modifications with ChIP-on-chip and ChIP-seq assays, where the H3K27me3 enriched sites, instead of sharply defined peaks, consist of broad, spread out signals, often referred to as blanketing [32].

We have demonstrated the power of a multiple validation work flow and the enrichR toolkit by performing an integrative analysis of a ChIP-on-chip dataset on estrogen receptor α. We extended the biological interpretation of the list of targets identified previously, and build evidence for a potential mechanism involving co-regulators in estrogen receptor α mediated gene regulation. All of this was done using existing data from online repositories and open source tools. By integrating the list of potential estrogen receptor α targets with gene expression data, we distinguished two classes of targets. The first class includes those targets that are induced upon binding of estrogen receptor α. Analysis of the promoter of these targets showed that ERE half sites are present. A gene ontology analysis on this set of targets indicated that these targets are mainly involved in processed linked to cell cycle and proliferation, in accordance with observations made previously. The second group of targets were those that upon binding of ER α are repressed. Motif analysis of these targets revealed ERE motifs more resembling the

canonical ERE motif, differing from the motif found in induced targets. Gene ontology enrichment analysis showed that these targets are mostly involved in metabolism and regulation of transcription. This suggests that the binding event of activated ER-α on the promoter is different for induced targets then it is for repressed targets. It is known from literature that endometrial proliferation is induced because activated ER-α recruits specific co-regulators at promoters of genes involved in the cell cycle regulation, which can be distinguished from co-regulators recruited/genes targeted by ER-α when its activation is not coupled to proliferation [6,7,8,13,26,27]. Our observations on the limited set presented here comply with these observations.

The enrichR software, a full description of all functions and required packages, in addition to the tables and script used for the analysis is freely available from http://www.bigcat.unimaas.nl/wiki/index.php/EnrichR.

Methods

enrichR

In general, the analysis of DNA enrichment microarray data comprises the following five steps:

- Data QC: Data quality control (QC) is no different from QC on other two-channel microarray technology and many free solutions have been developed and are readily available in R, of which the arrayQualityMetrics package is the a comprehensive and elegant solution [41].
- Data preprocessing: There are several key differences with other two-channel microarray technology, because many assumptions underlying the various approaches do not hold for DNA enrichment data [10].
- Enrichment finding: Enrichment finding is based on the assumption that it is possible to distinguish genuine enrichment from random effects within each sample. Most approaches are sliding window based and consist of at least two steps. The first is finding regions of enrichment, that is windows where the probes have a significantly higher signal distribution compared to the whole signal distribution of the microarray. The second is determining the false discovery rate of such an enrichment call, which in most cases determines how many times a window with a specific enrichment confidence level has that or a better confidence level through several data permutations. When the number of permutations is sufficient, this allows the estimation of a false discovery rate for each enriched window. The first step can use several tests to compare window distributions to the whole distribution, such as a two-sample Kolmogorov-Smirnov test or χ-square test on dichotomized data using a specific cutoff, as long as the number of probes in such a window is sufficient. In general, sufficient means around ten probes. Depending on the probe spacing, which averages somewhere between 50 and 100 bp for most modern microarrays, this means a window size of between 500 and 1000 bp. ACME uses the χ-square test approach for enrichment finding. A downside of this approach is that a cut-off is required, which the authors suggest be between 90th and 95^{th} quantile of the data. Although sufficient for exploratory data

analysis, this hampers the ability to detect smaller changes in enrichment, which is essential for DNA methylation microarray data. We therefore developed a similar approach using a two-sample Kolmogorov-Smirnov test that does not require such a cut-off. It compares the distribution of a set of probes in a window to the distribution of all probes, assessing if the set of probes in the window is drawn from a distribution that has a significantly higher mean than the distribution of all probes. This enrichment finding can easily be extended to multiple samples, such as replicates, taking for each window the probes from all samples and comparing them to the joint distribution of all samples.

- Differential enrichment analysis: Differential enrichment analysis can follow exactly the same approach, but instead of using the signal from one condition, we use the ratio of signals between conditions. As s second option, it can use the results from enrichment finding on one condition, and then compare these results afterwards. That is difficult however, as a score is needed for comparison that needs to enable statistical testing. We have developed two approaches for this. The first one uses the mean of the probes in a window as a measure of enrichment and compares them between conditions. In our experience 6 replicates per condition is sufficient for this approach. The second approach uses all the values of the probes in a window, performing an ANOVA between conditions using the values of all replicates. This is a more robust approach for smaller datasets, but decreases the power to call differences.
- Biological interpretation: With respect to biological interpretation, there is a clear need for sequence retrieval tools. We have created several functions to retrieve promoter sequences, sequences of enriched regions and a generic function to get sequences of any region of interest. Sequence retrieval is robust. If connection error, it retries. If something fails, it closes all open connections before quitting the function. The promoter retrieval tool requires Ensembl IDs, because of the use of Biomart. Because NimbleGen uses Entrez Gene identifiers in their promoter reports, we created a simple batch converter for Entrez Gene to Ensembl Gene ID.
  Additionally, we created a Gene Ontology enrichment function that only needs a set of Entrez Gene IDs of interest.

The enrichR toolkit comprises several functions for DNA enrichment data analysis (**table 2**). They are split into two portions: one set is generic and independent of platform and technology, the other set is specific for NimbleGen microarrays. The latter has functions for raw data and functions specialized for NimbleGen promoter report data, which is data already analyzed in-house by NimbleGen and the easiest starting point for an analysis. All functions require a minimum amount of input from the user and have suitable defaults for most common applications.

Function	Type	Description
getPromoterSequence	Generic	Get promoter sequences from Ensembl
entrez2synonym	Generic	Convert Entrez Gene IDs to other
myGetSequence	Generic	Retrieve genomic sequences from Ensembl
write.fasta	Generic	Write FASTA format sequence files
toEnsemblGeneId	Generic	Convert Entrez Gene IDs to Ensembl Gene ID
GOenrichmentAnalysis	Generic	Wrapper for easy GO enrichment analysis
cgiCheck	Generic	Check a sequence for the presence of a CpG island
nimbleGetPeakSequence	NimbleGen Promoterreport data	Gets peaks sequence
nimbleReadPromoterReport	NimbleGen Promoterreport data	Import NimbleGen promoter report
nimbleChromPlot	NimbleGen Promoterreport data	Create chromosome plot of enriched locations
nimbleFindCommonPeaks	NimbleGen Promoterreport data	Find common peaks in two samples
nimbleWriteGff	NimbleGen Promoterreport data	Creates GFF file
nimbleReadPos	NimbleGen raw data	Import NimbleGen probe postion file
nimbleReadRaw	NimbleGen raw data	Import NimbleGen raw intensity pair files
nimbleQC.Raw	NimbleGen raw data	Creates several QC plots to assess overall quality of the arrays
nimbleQC.Norm	NimbleGen raw data	Creates several QC plots to assess overall quality of the arrays
nimbleEnrichmentCalc	NimbleGen raw data	Calculated enrichment for each genomic region
nimbleDifferentialEnrichmentCalc	NimbleGen raw data	Calculate differential enrichment between conditions
nimbleCreateAnnotation	NimbleGen raw data	Using Biomart, automatically creates microarray annotation

Table 2: *An overview of the functions in the enrichR toolkit.*

Apart from the standard Bioconductor installation [34], several non-default packages are required. The most important ones are Ringo [35], RMySQL, RCurl [42], biomaRt [43], topGO [15] and GOsim [44].

The enrichR software, a full description of all functions and required packages is available from http://www.bigcat.unimaas.nl/wiki/index.php/EnrichR.

Datasets, analysis & script

The estrogen receptor α ChIP-on-chip dataset used for the demonstration was published previously [13]. The transcriptomics dataset is available from ArrayExpress (accession number E-GEOD-3834). The table and script used for the analysis are available from http://www.bigcat.unimaas.nl/wiki/index.php/EnrichR.

Author contributions

MA drafted the manuscript; MA and WA created, combined and tested the functional code; AR advised on the workflow and required tools for comprehensive biological interpretation and provided the data used throughout the manuscript; LE assisted in adapting and generalizing relevant parts of the code for future automation of the workflow; CE supervised the project.

References

1. Lazebnik l: Can a biologist fix a radio, or what I learned while studying apoptosis. Advances in gerontology 2003, 12:166-171.
2. Zheng M, Barrera LO, Ren B, Wu YN: ChIP-chip: Data, Model, and Analysis. Biometrics 2007, 63(3):787-796.
3. Mohn F, Weber M, Schübeler D, Roloff T-C: Methylated DNA Immunoprecipitation (MeDIP). Methods Mol Biol 2009, 507:55-64.
4. Esteller M. Cancer epigenomics: DNA methylomes and histone-modification maps. Nature Reviews 2007. (8): 286-298.
5. Ordway JM, Bedell JA, Citek RW, Nunberg A, Garrido A, Kendall R, Stevens JR, Cao D, Doerge RW, Korshunova Y et al: Comprehensive DNA methylation profiling in a human cancer genome identifies novel epigenetic targets. Carcinogenesis 2006, 27(12):2409-2423.
6. Carroll JS, et al. Genome-wide analysis of estrogen receptor binding sites. Nat Genet 2006. 38(11): 1289-97.
7. Zaret KS, Carroll JS: Pioneer transcription factors: establishing competence for gene expression. Genes & development 2011, 25(21):2227-2241.
8. Robinson JLL, Macarthur S, Ross-Innes CS, Tilley WD, Neal DE, Mills IG, Carroll JS: Androgen receptor driven transcription in molecular apocrine breast cancer is mediated by FoxA1. The EMBO journal 2011, 30(15):3019-3027.
9. Kubben N, et al. Genome-wide disruption of chromatin-lamina interactions by progerin. Chromosoma 2012.
10. Adriaens ME, et al. An evaluation of two-channel ChIP-on-chip and DNA methylation microarray normalization strategies. BMC Genomics 2012.
11. Scacheri PC, Crawford GE, Davis S: Statistics for ChIP-chip and DNase hypersensitivity experiments on NimbleGen microarrays. Methods Enzymol 2006, 411:270-282.
12. Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones SJM: FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics (Oxford, England) 2008, 24(15):1729-1730.
13. Romano A, Adriaens M, Kuenen S, Delvoux B, Dunselman G, Evelo C, Groothuis P: Identification of novel ER-alpha target genes in breast cancer cells: gene- and cell-selective co-regulator recruitment

at target promoters determines the response to 17beta-estradiol and tamoxifen. Mol Cell Endocrinol 2009, 314(1):90-100.

14. Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Holloway E et al: ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic acids research 2011, 39(Database issue):D1002-1004.
15. Alexa A, Rahnenführer J, Lengauer T: Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics (Oxford, England) 2006, 22(13):1600-1607.
16. Van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S, Conklin BR, Evelo C: Presenting and exploring biological pathways with PathVisio. BMC Bioinformatics 2008, 9(1):399.
17. Bembom O, Keles S, van der Laan MJ: Supervised detection of conserved motifs in DNA sequences with cosmo. Stat Appl Genet Mol Biol 2007, 6:Article8.
18. Hestand M, van Galen M, Villerius M, van Ommen G-J, den Dunnen J, t Hoen P: CORE_TF: a user-friendly interface to identify evolutionary conserved transcription factor binding sites in sets of co-regulated genes. BMC Bioinformatics 2008, 9(1):495.
19. Kuttippurathu L, Hsing M, Liu Y, Schmidt B, Maskell DL, Lee K, He A, Pu WT, Kong SW: CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments. Bioinformatics (Oxford, England) 2011, 27(5):715-717.
20. Shen L, Kondo Y, Guo Y, Zhang J, Zhang L, Ahmed S, Shu J, Chen X, Waterland RA, Issa J-PJ: Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. PLoS genetics 2007, 3(10):2023-2036.
21. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y et al: Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 2010, 466(7303):253-257.
22. Bock C, Paulsen M, Tierling S, Mikeska T, Lengauer T, Walter Jr: CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. PLoS genetics 2006, 2(3):e26.
23. McKay JA, Adriaens ME, Ford D, Relton CL, Evelo CTA, Mathers JC: Bioinformatic interrogation of expression array data to identify nutritionally regulated genes potentially modulated by DNA methylation. Genes Nutr 2008, 3(3-4):167-171.
24. Http://wiki.bigcat.unimaas.nl/ChIP_Course
25. Rao GG, Miller DS: Hormonal therapy in epithelial ovarian cancer. Expert Rev Anticancer Ther 2006, 6(1):43-47.
26. Laganière J, Deblois G, Lefebvre C, Bataille AR, Robert F, Giguère V: From the Cover: Location analysis of estrogen receptor alpha target promoters reveals that FOXA1 defines a domain of the estrogen response. Proc Natl Acad Sci U S A 2005, 102(33):11651-11656.

27. Jin VX, Leu YW, Liyanarachchi S, Sun H, Fan M, Nephew KP, Huang TH, Davuluri RV: Identifying estrogen receptor alpha target genes using integrated computational genomics and chromatin immunoprecipitation microarray. Nucleic Acids Res 2004, 32(22):6627-6635.
28. Lewis JS, Jordan VC: Selective estrogen receptor modulators (SERMs): mechanisms of anticarcinogenesis and drug resistance. Mutat Res 2005, 591(1-2):247-263.
29. Creighton CJ, Cordero KE, Larios JM, Miller RS, Johnson MD, Chinnaiyan AM, Lippman ME, Rae JM: Genes regulated by estrogen in breast tumor cells in vitro are similarly regulated in vivo in tumor xenografts and human breast tumors. Genome biology 2006, 7(4):R28.
30. Takai D, Jones PA: Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A 2002, 99(6):3740-3745.
31. Reiss DJ, Facciotti MT, Baliga NS: Model-based deconvolution of genome-wide DNA binding. Bioinformatics 2008, 24(3):396-403.
32. Young MD, Willson TA, Wakefield MJ, Trounson E, Hilton DJ, Blewitt ME, Oshlack A, Majewski IJ: ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. Nucleic acids research 2011.
33. Goecks J, Nekrutenko A, Taylor J, Galaxy T: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome biology 2010, 11(8):R86.
34. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J et al: Bioconductor: open software development for computational biology and bioinformatics. Genome biology 2004, 5(10):R80.
35. Toedling J, Sklyar O, Huber W: Ringo - an R/Bioconductor package for analyzing ChIP-chip readouts. BMC Bioinformatics 2007, 8(1):221.
36. Zacher B, Kuan PF, Tresch A: Starr: Simple Tiling ARRay analysis of Affymetrix ChIP-chip data. BMC bioinformatics 2010, 11:194.
37. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH: An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol 2008, 26(11):1293-1300.
38. Benoukraf T, Cauchy P, Fenouil R, Jeanniard A, Koch F, Jaeger Sb, Thieffry D, Imbert J, Andrau J-C, Spicuglia S et al: CoCAS: a ChIP-on-chip analysis suite. Bioinformatics (Oxford, England) 2009, 25(7):954-955.
39. Pelizzola M, Koga Y, Urban AE, Krauthammer M, Weissman S, Halaban R, Molinaro AM: MEDME: an experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment. Genome research 2008, 18(10):1652-1659.
40. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Gräf S, Johnson N, Herrero J, Tomazou EM et al: A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat Biotechnol 2008, 26(7):779-785.

41. Kauffmann A, Gentleman R, Huber W: arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. Bioinformatics (Oxford, England) 2009, 25(3):415-416.
42. Lang DT: R as a Web Client – the RCurl package. Journal of Statistical Software 2007, VV(II).
43. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W: BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics (Oxford, England) 2005, 21(16):3439-3440.
44. Fröhlich H, Speer N, Poustka A, Beissbarth T: GOSim--an R-package for computation of information theoretic GO similarities between terms and gene products. BMC bioinformatics 2007, 8:166.

Chapter 3

An evaluation of two-channel ChIP-on-chip and DNA methylation microarray normalization strategies

Michiel E. Adriaens[1,2,*], Magali Jaillard[1,*], Lars M.T. Eijssen[1], Claus-Dieter Mayer[3] and Chris T.A. Evelo[1,2]

** Equal contribution.*

[1]*Department of Bioinformatics - BiGCaT, Maastricht University, Maastricht, The Netherlands*
[2]*Netherlands Consortium for Systems Biology (NCSB), University of Amsterdam, The Netherlands*
[3]*Department of Biomathematics and Statistics Scotland, University of Aberdeen, Rowett Institute of Nutrition and Health, Aberdeen, UK*

Keywords: *ChIP-on-chip, MeDIP-on-chip, DNA methylation, two-channel microarray, normalization.*

Publication: Adriaens, M.E., M. Jaillard-Dancette, L. Eijssen, C. Mayer and C.T.A. Evelo (2012). "Normalization strategies for two-channel ChIP-on-chip and DNA methylation microarray technologies." BMC Genomics.

Abstract

The combination of chromatin immunoprecipitation with two-channel microarray technology enables genome-wide mapping of binding sites of DNA-interacting proteins (ChIP-on-chip) or sites with methylated CpG di-nucleotides (DNA methylation microarray). These powerful tools are the gateway to understanding gene transcription regulation. Since the goals of such studies, the sample preparation procedures, the microarray content and study design are all different from transcriptomics microarrays, the data pre-processing strategies traditionally applied to transcriptomics microarrays may not be appropriate. Particularly, the main challenge of the normalization of "regulation microarrays" is (i) to make the data of individual microarrays quantitatively comparable and (ii) to keep the signals of the enriched probes, representing DNA sequences from the precipitate, as distinguishable as possible from the signals of the un-enriched probes, representing DNA sequences largely absent from the precipitate. We compare several widely used normalization approaches (VSN, LOWESS, quantile, T-quantile, Tukey's biweight scaling, Peng's method) applied to a selection of regulation microarray datasets, ranging from DNA methylation to transcription factor binding and histone modification studies. Through comparison of the data distributions of control probes and gene promoter probes before and after normalization, and assessment of the power to identify known enriched genomic regions after normalization, we demonstrate that there are clear differences in performance between normalization procedures. T-quantile normalization applied separately on the channels and Tukey's biweight scaling outperform other methods in terms of the conservation of enriched and un-enriched signal separation, as well as in identification of genomic regions known to be enriched. T-quantile normalization is preferable as it additionally improves comparability between microarrays. In contrast, popular normalization approaches like quantile, LOWESS, Peng's method and VSN normalization alter the data distributions of regulation microarrays to such an extent that using these approaches will impact the reliability of the downstream analysis substantially.

Note: high-resolution figures freely available online at http://www.biomedcentral.com/1471-2164/13/42

Background

For over a decade, two-channel transcriptomics microarrays have provided a powerful approach to study genome-wide gene expression events. Now, continued development of two-channel microarray technology has enabled extending our experimentation to the next level: regulation of gene transcription. One of the most popular techniques in this field combines chromatin immunoprecipitation (ChIP) assays with two-channel microarray technology (ChIP-on-chip [1]). ChIP-on-chip studies are used to detect any protein-DNA interaction genome-wide, such as transcription factor binding, but also epigenetic events such as histone modifications, as long as a suitable antibody is available. The same approach is used to detect DNA methylation, by using either an antibody that interacts with methyl-CpG-binding domain (MBD) proteins bound to methylated CpG dinucleotides (MBD-ChIP assay), or an antibody that interacts with methylated CpG dinucleotides directly (methylated DNA immunoprecipitation (MeDIP) assay [2]).

Even though newer technologies such as ChIP-sequencing (ChIP-seq) are on the rise, two-channel microarrays still present a valuable approach to understanding gene transcription regulation events, and during the last decade have opened opportunities to identify novel targets and markers in complex diseases such as cancer [3,4,5], heart failure [6] and diet-related disorders [7], and psychiatric disorders such as depression, schizophrenia and addiction [8]. Since the main appliance of this technology at the time being is gene transcription regulation studies – transcription factor and co-regulator binding, DNA methylation, and histone modifications – the term 'regulation microarrays' will be used for brevity henceforth.

The design and the experimental approach for regulation microarrays are very different from the more extensively studied transcriptomics microarrays, which has implications for data pre-processing procedures. The key difference is that in transcriptomics microarrays both channels contain amplified transcript samples, usually corresponding to two different experimental conditions, while in regulation microarrays the channels comprise an experimental sample and a reference sample. The cyanine 3 (Cy3), or green, channel of regulation microarrays generally contains the total DNA sample that gives the reference baseline signal, and the cyanine 5 (Cy5), or red, channel contains an experimentally enriched DNA sample, extracted using a specific antibody binding to a DNA-interacting protein (ChIP) or directly to methylated CpGs on the DNA (MeDIP). Hence, while the log-ratio between the channel signals represents the differential expression between two conditions in transcriptomics studies, for regulation microarrays it is used as a measure of enrichment: the higher the log-ratio of a probe or set of tiling probes, the higher the likelihood that the corresponding region in the genome has a high level of methylation or is targeted by a DNA-interacting protein.

Another important assumption in regulation microarrays is that a DNA-interacting protein is either bound or not bound (for ChIP) and that a target sequence is either methylated or not (for MeDIP). Regardless, depending on binding affinity, mean time of residence and other factors, the fraction of cells with bound protein or a particular methylation status is not an all-or-nothing condition, especially in heterogeneous tissues. Combined with the characteristics of the data distribution surrounding a site of interest (**figure 1**) and probe effects [9], this produces a continuous log-ratio distribution. However, the characteristics of the samples hybridized to the channels force a dichotomy upon the log-ratio distribution, which is comprised of two components (**figure 2**) commonly referred to as an enriched and an un-enriched component [10]. The enriched component corresponds to the probes to which the experimental DNA has hybridized and the un-enriched component to the probes whose targets are largely absent from the experimental DNA sample. Hence, contrary to transcriptomics microarray data, where low log-ratio values are meaningful as long as the differences between conditions are statistically significant, when interpreting ChIP-on-chip and DNA methylation microarray data, the upper quantile is of most interest, as it generally comprises mostly enriched probes. Based on this assumption, enrichment finding algorithms like ACME [11], will test if a set of tiling probes is significantly more likely to be a sampling of this upper quantile than of the rest of the

data, assuming that if this is the case, this set of tiling probes corresponds to a protein binding site or methylated region. A better separation between the enriched and un-enriched components hence increases the power to identify enriched regions. Thus, a crucial aspect in regulation studies is that any separation between the enriched and un-enriched components present in the data before normalization, should be kept afterwards. Apart from conserving this separation, other aspects need to be taken into account when normalizing regulation microarray data.

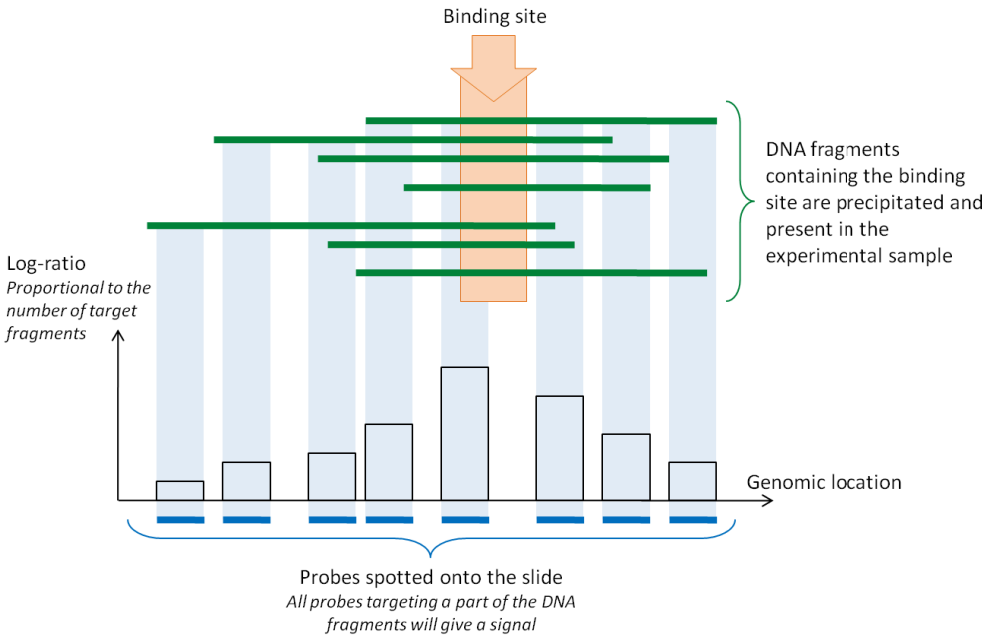


Figure 1: *The birth of an enrichment signal around a binding site (ChIP-on-chip). Since DNA fragmentation through sonication can be modeled as a Poisson process [1], the DNA fragment length distribution follows a Poisson distribution and adjacent probes on the genome have a correlated log-ratio, resulting in the hybridization pattern shown here. Each blue column represents a probe hybridization site. Black-outlined bars represent their log-ratio. Green lines are sonicated immuno-precipitated DNA fragments corresponding to the binding site.*

Normalization is a process that is applied at multiple levels connected to spatial [12], probe [13,14,15] and dye or intensity dependent biases [16]. Additionally, differences in print quality, differences in ambient conditions when the plates were processed or changes in the scanner settings can cause scaling differences between microarrays. Most of the assumptions underlying the process of correcting for these biases are identical for transcriptomics microarrays and regulation microarrays. The exception is the correction for intensity dependent bias, for which the most common approaches in use for transcriptomics microarrays are LOWESS normalization [12,17,18] and quantile normalization [19]. Both methods are based on the assumption that the majority of probe signals are unchanged between channels and

microarrays, which generally holds for transcriptomics studies [12,20,18]. In regulation studies however, this assumption does not hold since the samples comprising the two channels differ to a large extent.

Based on these observations, the main challenge of the normalization of regulation microarrays is (i) to make the signals of individual microarrays quantitatively comparable and (ii) to retain the separation between the enriched and un-enriched components present in the data. Programs like CoCAS [21] offer a range of normalization methods for regulation microarrays, including quantile normalization [19] and variance stabilizing normalization [22], and R/Bioconductor [23] offers many more popular choices, which may not all be suitable for this challenge. Hence, we here assess the efficacy in removing technical biases and in preservation of the separation between the enriched and un-enriched components, for six two-channel microarray normalization methods (VSN [22], LOWESS [12,16], quantile [19], T-quantile [19], Tukey's biweight scaling, Peng's method [24]) applied to five published ChIP-on-chip and MeDIP-on-chip datasets on the NimbleGen platform.

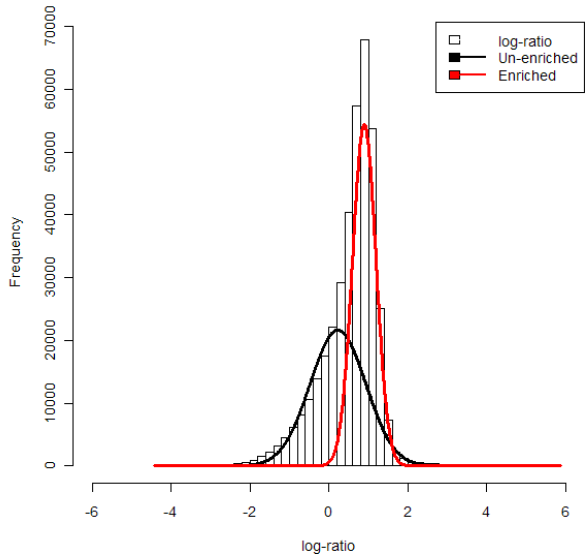


Figure 2: *An example of a two-component distribution fitted on ChIP-on-chip data of dataset #1 (see Methods section for dataset description and numbering).*

Results

To determine the efficacy in correcting for technical biases and improving comparability between microarrays, quality control and bias assessment was performed on all datasets before and after normalization for each of the six normalization approaches. Complete results are available from http://www.bigcat.unimaas.nl/userfiles/adriaens/arrayQualityMetrics/. In all datasets scaling effects between microarrays and intensity dependent bias within microarrays are present, visible from the microarray data distributions. All tested normalization methods are able to correct for the observed

biases, where from a technical standpoint, normalization approaches that normalize channels together (VSN, LOWESS, Peng's method, quantile) equalize the data distributions to a larger extent than normalization approaches that normalize the channels separately (T-quantile, Tukey's biweight scaling). In the latter category, T-quantile normalization enhances overall comparability to a larger extent than Tukey's biweight scaling.

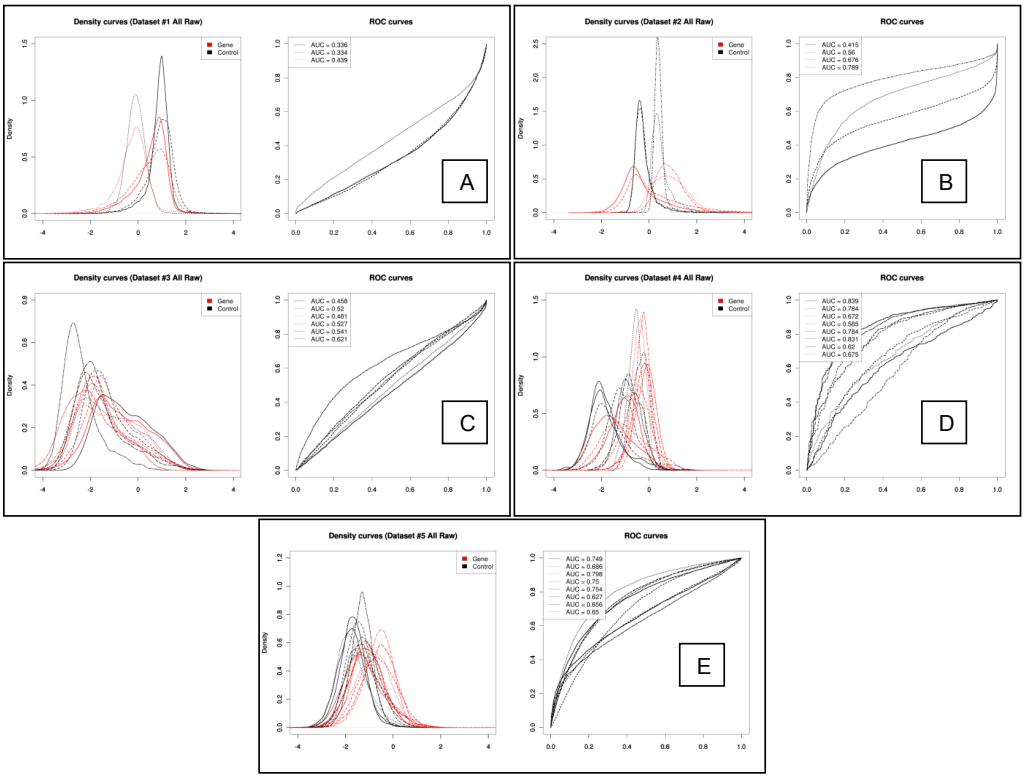


Figure 3: *Density distributions of the control probes and gene promoter probes of the raw log-ratio data of all individual microarrays and corresponding ROC curves for dataset #1 (a), dataset #2 (b), dataset #3 (c), dataset #4 (d) and dataset #5 (e). AUC values of each ROC curve are reported in the legend.*

To evaluate the separation between the enriched and un-enriched components, the gene promoter probe and the negative control probe log-ratio distributions were assessed using ROC curves before and after normalization with each of the six normalization approaches (**figure 3**). The raw data from dataset #1 (see **Methods** section for dataset details and numbering) shows largely overlapping control probe and gene promoter probe distributions (**figure 3a**). Between individual microarrays, the distributions show larger differences, also resulting in more variation in both the area under the curve (AUC) as well as the

shape of the ROC curves, indicating that comparability between microarrays is hindered by lack of normalization.

The results of the combined data of the individual microarrays from the six normalization approaches (VSN, LOWESS, quantile, T-quantile, Tukey's biweight scaling, Peng's method) show equal performance of all approaches for dataset #1 (**figure 4a**), resulting in ROC curves with similar shape and comparable AUC values. Based on the AUC values, separation between components is best when using Peng's method.

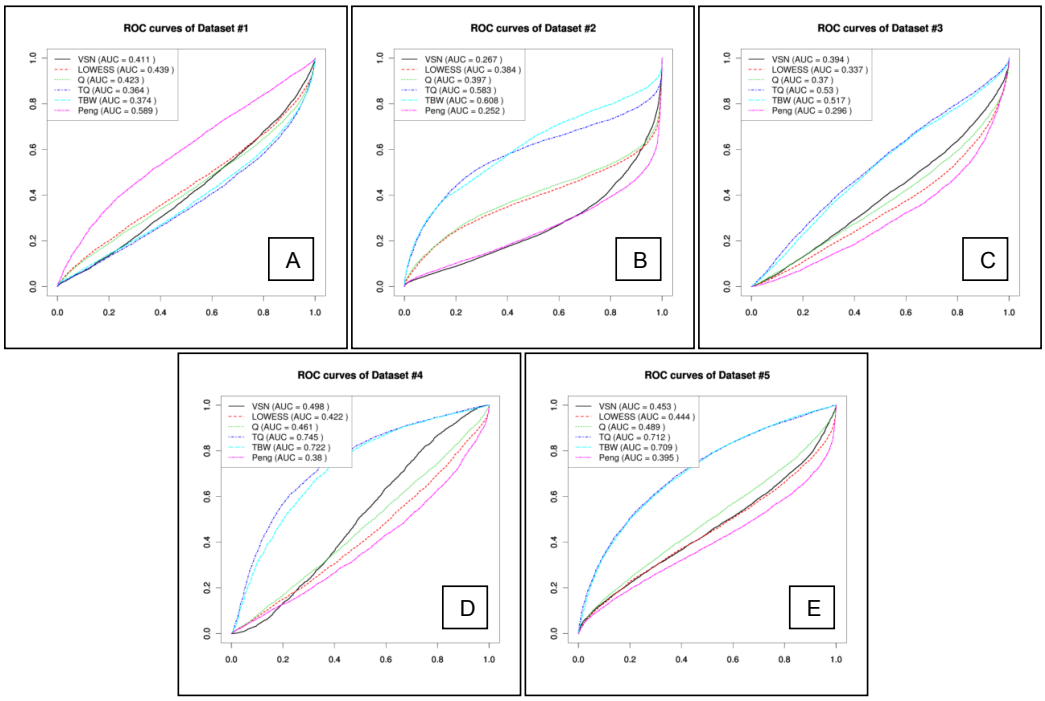


Figure 4: *ROC curves of the control probe and gene promoter distributions of the combined log-ratio data, for each normalization approach of dataset #1 (a), dataset #2 (b), dataset #3 (c), dataset #4 (d) and dataset #5 (e). AUC values are reported in the legend. TBW = Tukey's biweight scaling, Q = quantile normalization, TQ = T-quantile normalization.*

Dataset #2, the second ChIP-on-chip dataset gives different results (**figure 3b and 4b**). Separation between components is preserved best when using T-quantile or Tukey's biweight scaling normalization (**figure 4b**). The other approaches, including Peng's method, alter the ranking of probes resulting in the control probe and gene promoter probe distributions becoming superimposed. VSN normalization appears to scale the distributions, enforcing a larger spread compared to the data acquired through the other normalization approaches (**figure 5**).

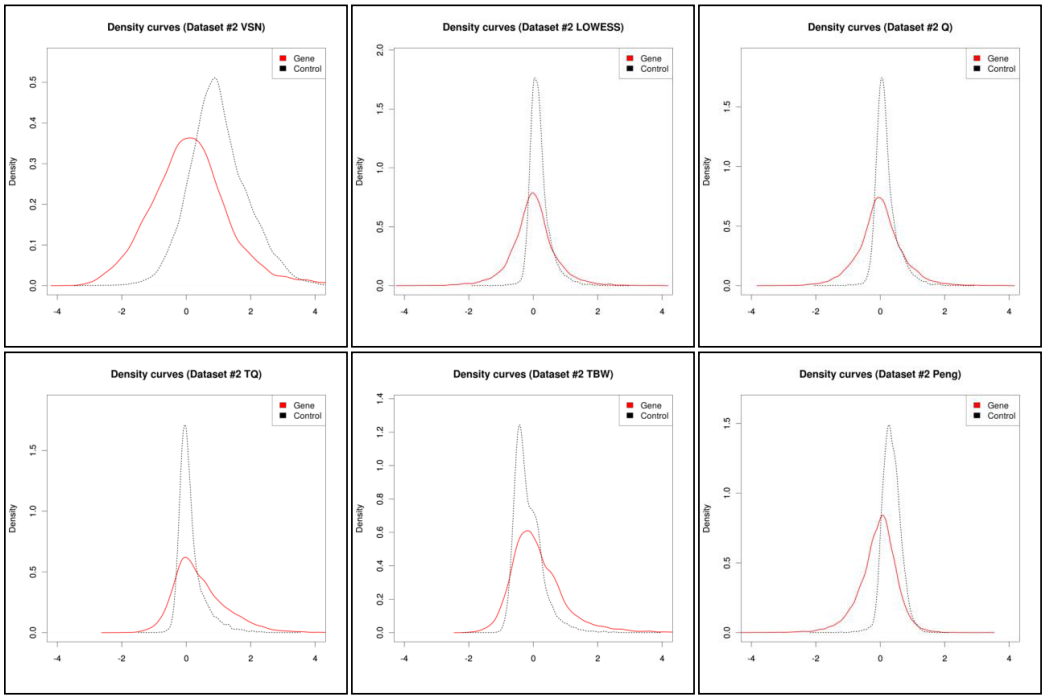


Figure 5: *Density distributions of the control probes and gene promoter probes of the normalized combined log-ratio data of dataset #2 (ChIP-on-chip). Results are shows for (from left to right, top to bottom) VSN, LOWESS, quantile (Q) , T-quantile (TQ), Tukey's biweight scaling (TBW), Peng's method.*

Tukey's biweight scaling and T-quantile normalization appear to perform comparably with respect to conserving the component separation. Tukey's biweight scaling adjusts the log-ratio data with a scaling factor for each microarray in the dataset individually, which means that the ROC curves will be identical to those of the raw data, and that the distributions will be the same as those before normalization save for a shift. This may explain the variability observed in the individual ROC curves and AUC values of the Tukey's biweight scaling normalized data. T-quantile normalization reduces the variability between the data distributions of individual microarrays, resulting in ROC curves that are more comparable in both shape and AUC (**figure 6**).

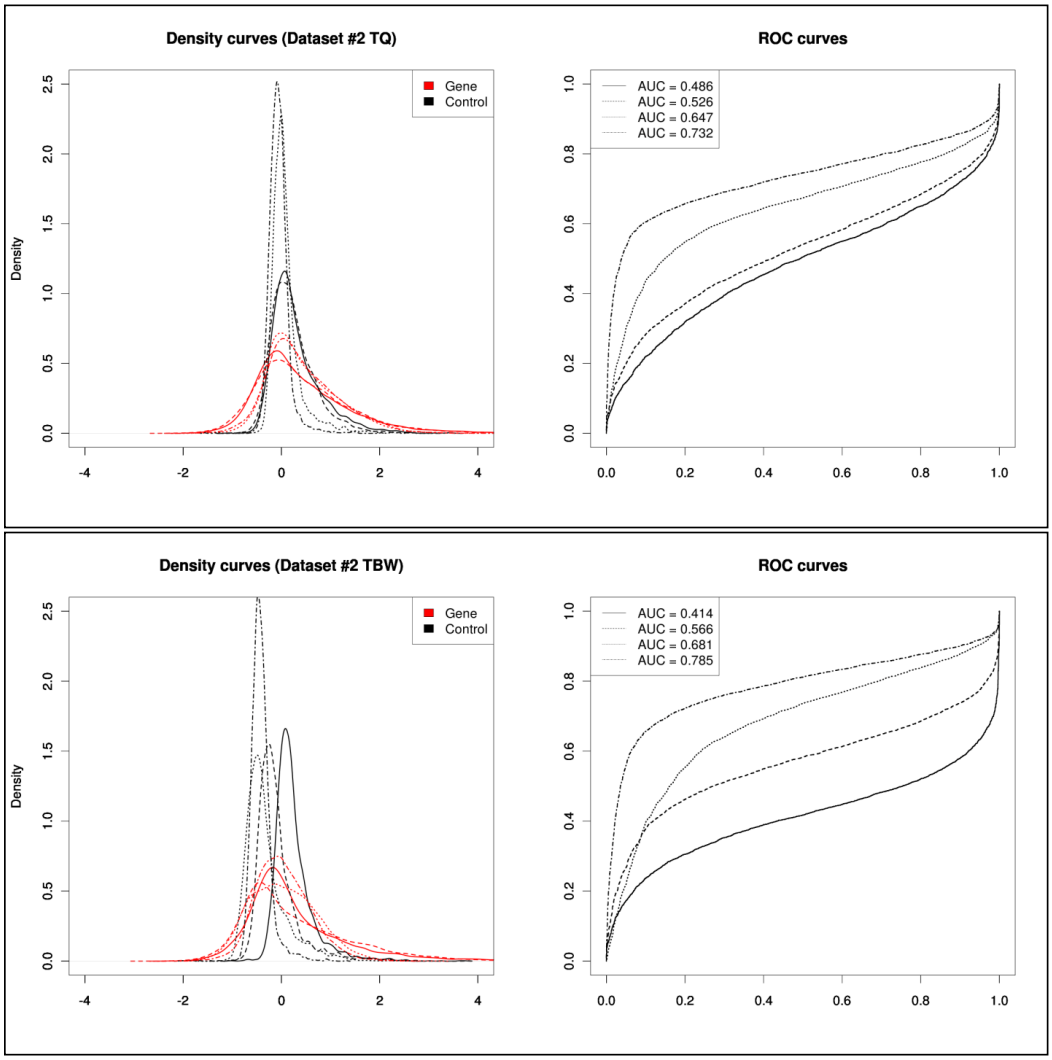


Figure 6: *Density distributions of the control probes and gene promoter probes of the normalized log-ratio data of each individual microarray and corresponding ROC curves of dataset #2 (ChIP-on-chip). Top: Results for T-quantile (TQ) normalized data. Bottom: Results for Tukey's biweight scaling (TBW) normalized data. AUC values of each ROC curve are reported in the legend.*

The results of the MeDIP-on-chip datasets support the conclusions reached for the ChIP-on-chip datasets: separation of the components present before normalization (**figure 3c, 3d and 3e**) are preserved best with T-quantile and Tukey's biweight scaling approaches (**figure 4c**). LOWESS, quantile, VSN and Peng's normalization alter the distributions and eradicate the separation. In dataset #3, the differences between the normalization approaches is less striking, illustrated by similarly shaped ROC

curves and AUC values (**figure 4c**). Dataset #4 shows a larger heterogeneity between individual microarrays than both dataset #3 and #5. For both dataset #4 and #5 Tukey's biweight scaling and T-quantile normalization produce higher AUC values for these approaches (**figure 4d and 4e**). Both methods appear to perform comparably with respect to conserving the component separation. However, as in dataset #2, the differences between both approaches are highlighted by the distributions of the individual microarrays: Tukey's biweight scaling adjusts each microarray individually, whereas T-quantile normalization is applied between microarrays. T-quantile normalization thereby results in ROC curves with less variation in shape and AUC (**figure 7**) than those of the raw data and the Tukey's biweight scaling normalized data.

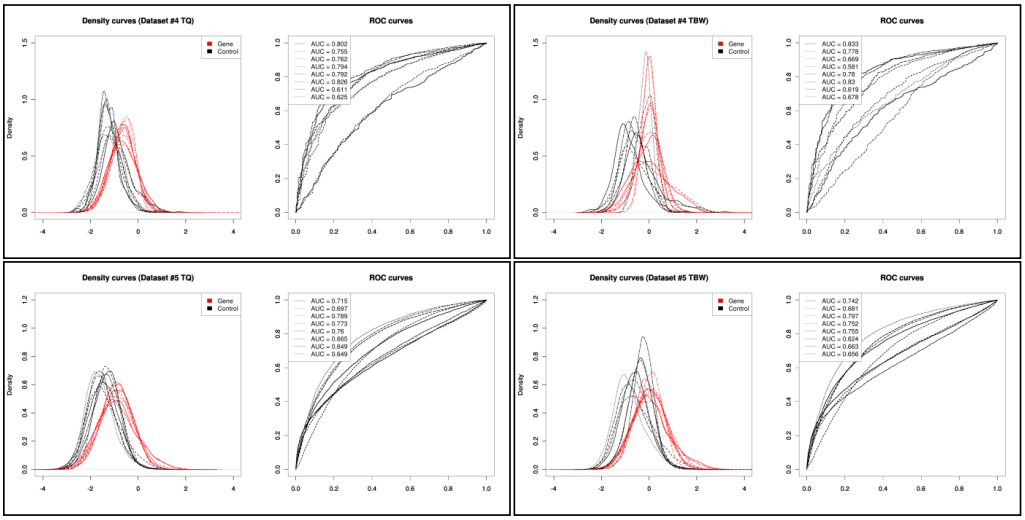


Figure 7: *Density distributions for the control probes and gene promoter probes of the normalized log-ratio data of each individual microarray and corresponding ROC curves of dataset #4 and #5. Top left: Results for T-quantile (TQ) normalized data of dataset #4. Top right: Results for Tukey's biweight scaling (TBW) normalized data of dataset #4. Bottom left: Results for T-quantile (TQ) normalized data of dataset #5. Bottom right: Results for Tukey's biweight scaling (TBW) normalized data of dataset #5. AUC values of each ROC curve are reported in the legend.*

Any appropriate normalization method should preserve the biological information present in the raw data. Assessing the distributions of the negative control probes and the gene promoter probes is a global indicator of this conservation of biological information. In addition, three datasets with suitable positive controls were used to assess the impact of the normalization approaches on the power to identify significant enrichment for specific genomic regions. ACME [11] was used for all enrichment calculations. For dataset #1, 33 validated ER-a targets were used as positive controls [25,26]. The results are reported in **table 1** for all normalization approaches and for several enrichment p-value cut-offs (0.05, 0.10, 0.20

and 0.50). T-quantile and quantile normalization in general result in identification of more targets at each cut-off.

	Number of ER-a targets found (out of 33)			
Normalization approach	Enrichment p-value < 0.05	< 0.10	< 0.20	< 0.50
VSN	7	8	11	23
LOWESS	6	7	10	25
Quantile	8	9	12	25
T-quantile	9	10	14	24
Tukey's biweight scaling	7	8	11	23
Peng's method	5	5	9	23

Table 1: *Number of validated estrogen receptor α targets found significantly enriched in the estrogen receptor α ChIP-on-chip dataset (dataset #1). This table contains for each of the tested normalization approaches the number of validated estrogen receptor α targets [25] found significantly enriched in the estrogen receptor α ChIP-on-chip dataset (dataset #1). Results for four enrichment p-value cut-offs are given (0.05, 0.10, 0.20 and 0.50).*

For dataset #3 enrichment of the HOXA group of developmental genes was calculated. HOXA genes are located in a cluster on chromosome 7 and are known to be switched off and moderately to highly methylated in most tissues [27]. The negative ^{10}log-transformed enrichment p-values plotted along the HOXA region are shown in **figure 8** (**top**). Using Tukey's biweight scaling or T-quantile normalization results in identification of several enriched loci, most of which are moderately methylated. Less loci are found when using VSN, quantile or LOWESS normalization. Peng's method results in identification of only a few loci with moderate enrichment.

For dataset #4 enrichment was determined for the Dlk1-Gtl2 cluster on chromosome 12, a region reported in the original results [28] to be highly enriched. For all normalization approaches in dataset #4 the same area in this region is identified as very highly enriched (**figure 8 bottom**).

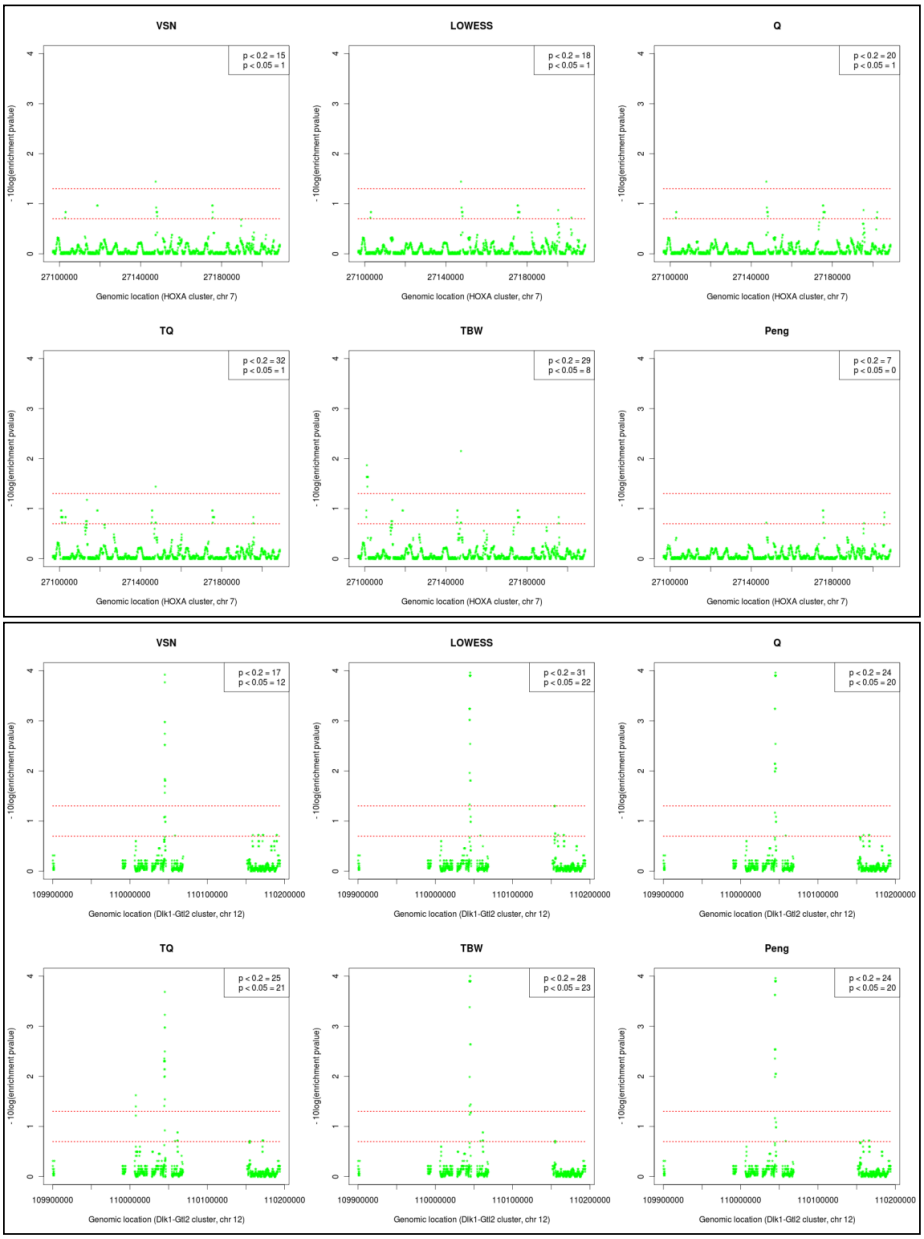


Figure 8: *Genome plots of negative ^{10}log-transformed enrichment p-values, for the HOXA cluster on human chromosome 7 (top) and the Dlk1-Gtl2 cluster on mouse chromosome 12 (bottom). Red vertical lines are given at values corresponding to p-values of 0.05 (top line) and 0.20 (bottom line). Regions with values above the top line are highly enriched, while values between the lines are a sign of moderate enrichment. The total number of identified enriched regions are reported in the legend. TBW = Tukey's biweight scaling, Q = quantile normalization, TQ = T-quantile normalization.*

Discussion

Two-channel transcriptomics and regulation microarrays should not be pre-processed in the same manner. Appropriate normalization strategies for regulation microarrays are characterized by their ability to retain the separation between the enriched and un-enriched components present in the data whilst enhancing comparability between microarrays. Six normalization methods were tested by (i) assessing the separation between the control probe and gene promoter probe distributions before and after normalization using ROC curves and (ii) by verifying whether known enriched genes and regions could be identified as such after normalization. We have shown that the result of each approach depends heavily on the situation before normalization, specifically the amount of enriched and un-enriched probes and the separation between the corresponding components in the raw data. These two characteristics are different for each experiment, depending largely on the biological system studied and the applied assay.

In the ChIP-on-chip datasets used here, the distributions of the control probes and gene promoter probes overlap to a large extent before and after normalization. This may be explained by the small proportion of the genome generally covered by the potential binding sites of a DNA-interacting protein and the resulting small contribution of the enriched component. Hence in general, the lower the amount of binding sites, the more similar the control probe and gene promoter distributions, and the more comparable the performance of the normalization approaches, based on ROC curves of both distributions before and after normalization. However, in some cases VSN can cause a sizeable rescaling of the distributions, and to a spurious control probe distribution with a higher mean and spread than the gene promoter probe distribution. This renders gene promoter probes in the upper quantile of the data indistinguishable from random data, strongly impacting the biological interpretation.

In DNA methylation microarrays the amount of enriched probes and un-enriched probes is of the same order, since in general the proportion of methylated CpG di-nucleotides in a genome is substantial. We have shown that for such microarrays, the choice for a normalization procedure will be crucial for the downstream analysis. All three MeDIP-on-chip datasets show a large degree of separation between the gene promoter and control probe distributions. The separation is lost when using normalization methods that normalize channels together, such as VSN, LOWESS, Peng's method and quantile normalization. Using LOWESS approaches on MeDIP-on-chip data has been reported elsewhere to result in increased bias, because the underlying assumption that the log-ratio should be independent of the average individual channel signals does not hold for this type of data. DNA methylation levels are related to CpG and GC density, while signal intensity is also known to be influenced by GC content. Forcing the log-ratio to be independent of the average signal intensity using LOWESS normalization thus introduces bias instead of removing it [9].

T-quantile normalization, applied separately on the channels, and Tukey's biweight scaling are the only approaches that are able to preserve the component separation in all example datasets. In dataset #1, individual microarrays already showed comparable distributions before normalization; hence for this dataset, Tukey's biweight scaling would be sufficient. In contrast, dataset #4 for example showed a large heterogeneity between individual microarrays, in which case between-microarray normalization is better

suited to improve the overall comparability and enable quantitative data comparison. This can be achieved either by doing an additional normalization step after scaling, but ideally by using a between-microarray normalization approach from the beginning, such as applying T-quantile normalization as demonstrated here.

In regulation microarrays the sequence content of the input DNA sample and the experimental DNA sample always differs to a large extent. There are also instances for transcriptomics microarrays, such as dedicated microarrays designed for a specific biological context, where the assumption that the majority of genes are not differentially expressed does not hold, hence requiring adapted normalization strategies. Most of these strategies involve the use of invariant genes, either present on the slide [29,30] or determined from the data [31]. Selecting invariant probes in ChIP-on-chip and DNA methylation data is difficult however, even when selecting the control probes used in the analysis presented here, because this would implicate a normalization based on un-enriched probes. Since the sequences meant to hybridize to these probes are largely absent from the experimental sample, they essentially measure background noise in the channel containing the experimental sample. Variation in log-ratio values of these un-enriched probes between microarrays therefore reflects methodological effects rather than biology, which compromises their usability. To avoid the use of invariant genes in transcriptomics microarrays, a three-component mixture model has been proposed [32]. The normalization parameters are estimated independently in the groups of up-regulated, down-regulated and unchanged genes and normalized separately. Such a model in adapted form can be fitted on regulation microarray data and used conjointly with enrichment finding. It has been shown that for DNA methylation studies using specific reference samples, such as a fully methylated total DNA sample, it is possible to make robust estimates for methylation percentages when using such a model [9, 33, 34].
The research described herein is limited to the normalization of replicate microarrays. In many cases however, a study will consist of multiple conditions, such as different tissues, or treatment and control samples as demonstrated in dataset #1. In these cases, the experimental DNA samples may differ to a large extent between treatment and control groups, warranting application of normalization to each condition separately. However, when only a relatively small amount of loci is expected to be differentially enriched and the total amount of enrichment can be assumed constant between conditions, normalization approaches applied to the dataset as a whole are more appropriate. This holds for experiments such as DNA methylation studies on the same tissue treated with a micronutrient [35], where only a projected limited amount of important regulatory regions with substantially altered levels of methylation is of interest.

The results of known targets and enriched regions show consistent differences between the various normalization approaches. When looking at the Dlk1-Gtl2 cluster for the DNA methylation data of dataset #4, a region reported to be highly enriched in the original findings, it is clear that such highly enriched

regions will be identified as such regardless of the chosen normalization approach. This is not the case when studying moderately enriched regions, as illustrated by the results of the HOXA cluster in dataset #3, where the degree to which this region is identified as being enriched depends strongly on the applied normalization approach. Overall, T-quantile normalization and Tukey's biweight scaling again give the best results. A potential cause of the observed difference between the tested normalization approaches is observed in the results on global level: the ranking of probes changes when using some normalization approaches, increasing the likelihood of un-enriched probes being spread over the whole dynamic range of the enriched probe distribution. Ultimately, such changes in the ranking can be destructive on the power to call differences in methylation or protein binding. Also, enrichment finding algorithms [11] as used for these results, will test if a group of tiling probes is significantly more likely to be part of the upper quantile than of the rest of the data distribution, assuming that if this is the case, this group of tiling probes shows significant enrichment and thus corresponds to a binding site or methylated region. This upper quantile can be defined for each microarray individually after normalization. Hence, it is not the values themselves, but the rank in the data distribution which is biologically relevant. Considering this, within channel and treatment normalization approaches do not only enable a more robust data interpretation, but since for many applications the individual values themselves do not need to be comparable, they are also sufficient.

Conclusion

The main issue of ChIP-on-chip and DNA methylation microarray normalization is to enhance comparability between microarrays, while keeping the separation between the enriched and un-enriched components present in the data. Within-channel approaches give the best performance, with enhanced comparability between individual microarrays for approaches that also normalize between microarrays. More specifically, quantile, LOWESS, Peng's and VSN normalization alter the signal distributions to such an extent that it will impact the reliability of the downstream analysis substantially. Better results are obtained with T-quantile normalization applied separately on the channels or Tukey's biweight scaling. For all datasets tested, these two methods consistently outperform the other tested methods in conservation of separation between the enriched and un-enriched distributions, as well as in identification of genomic regions known to be enriched. The T-quantile approach is preferable because it additionally yields enhanced comparability between microarrays.

Methods

ChIP-on-chip and DNA methylation microarray dataset selection

Five published datasets were selected from ArrayExpress. Selection criteria were set to select several assay types (MeDIP and ChIP), several species (human and mouse) and cover several research fields. Due to the selection criteria, all datasets were chosen from the same microarray manufacturer, NimbleGen (**table 2**).

Dataset number	#1	#2	#3	#4	#5
ArrayExpress ID	not registered	E-TABM-529	E-GEOD-17581	E-GEOD-24286	E-GEOD-22831
Assay type	ChIP-on-chip	ChIP-on-chip	MeDIP-on-chip	MeDIP-on-chip	MeDIP-on-chip
Microarray ID	NimbleGen Human HGS17 minimal promoter	NimbleGen Mouse Tiling 2006-07-17 MM8Tiling Set17	NimbleGen Homo sapiens 385K CGH array	NimbleGen mouse 385K Refseq and miRNA promoter tiling (2-array set)	NimbleGen Nimblegen HD2 MM8 promoter deluxe array
Species	Human	Mouse	Human	Mouse	Mouse
Investigation	Identification of ER-α target genes in breast cancer cells	Identification of histone modification profiles in WT and Kcnq1ot1	Methylome analysis of congenital ectopic thyroids	Mecp2-dependent regulation of MicroRNAs in Rett Syndrome	DNA methylation analysis in E3.5 blastocysts, E6.5 epiblasts and E9.5 whole embryos
No of microarrays	8	11	6	8	11
Microarray content	· 3 stimulated by 17beta-estradiol · 1 pool of the 3 stimulated · 3 untreated · 1 pool of the 3 untreated	· 2 Kcnq1ot1 · 9 wild type · Tissues are placenta or liver	· 3 orthotopic thyroid · 3 congenital ectopic thyroid	· 2 KO using Mecp2 · 4 wild type using Mecp2 · 2 wild type using 5-methylcytosine	· 2 E3.5 blastocysts · 3 E6.5 epiblasts · 3 E9.5 whole embryos · 3 Control pooled unamplified MeDIPs in E9.5 embryos
Microarrays used for this study	3: 2 stimulated + the pool of stimulated	4: H3K27me3 in wild-type placenta	6: all	8: all	8: all except the pooled controls
Data publication date	Article publication: 15/01/2010 PMID: 19698761	08/01/2008	27/10/2010	30/09/2010	01/11/2010

Table 2: *Technical information on the ChIP-on-chip and MeDIP-on-chip datasets used for the normalization approach comparison. This table contains all relevant the technical information of the ChIP-on-chip and MeDIP-on-chip datasets used for the normalization approach comparison, including the dataset number as used herein, the ArrayExpress ID, the assay type, the microarray ID, species, the total number of microarrays in the dataset, the experimental content of the microarrays, a specification of the subset of microarrays used for the analyses, and the publication date of the dataset on ArrayExpress.*

Sub-selections of microarrays and experimental groups were made to keep only the microarrays of sufficient quality and homogeneous replicate groups of sufficiently large size. In dataset #1, one microarray of the 17beta-estradiol stimulated group was removed because the red channel was saturated, as reported previously [26]. Instead the microarray containing a pool of stimulated samples

was included. The microarrays corresponding to the untreated group were left out of the analysis. In dataset #2, only the microarrays containing the wild-type placenta H3K27me3 samples were chosen. All the microarrays from dataset #3 and #4 were used. In dataset #5 all microarrays were used, except for three containing pooled samples.

Quality control and bias assessment of the raw and normalized data was performed using the arrayQualityMetrics package [36]. Individual reports are available online at http://www.bigcat.unimaas.nl/userfiles/adriaens/arrayQualityMetrics/.

Removing technical biases through normalization

Microarray data is subject to multiple sources of variation. The goal of normalization is to remove all technical biases from the microarray data, while retaining the biological variation. There are many normalization procedures available for two-channel microarray data, but the choice for a specific procedure has to be fuelled by the characteristics of the dataset: (i) the procedure should correct all the systematic biases in the dataset diagnosed during the quality control process and (ii) the underlying assumptions of the particular method must be met. In regulation studies, there is the additional goal to retain the separation between the enriched and un-enriched components of the log-ratio distribution.

To illustrate this, data from five human and mouse ChIP-on-chip and MeDIP-on-chip datasets were normalized using six different methods: (i) LOWESS normalization [12,16] applied on each microarray individually, which assumes the log-ratio distribution is a normal distribution centered around zero; (ii) Quantile normalization [19] applied between microarrays, which equalizes the intensity distributions of all channels – green and red – together; (iii) Variance stabilizing normalization (VSN) [22], which is applied between microarrays and between channels; (iv) T-quantile normalization [19], which allows for quantile normalization of the data in subgroups and here is applied to normalize the red and green channels separately; (v) Tukey's biweight scaling, which scales the log-ratio distribution of each microarray individually using a robust Tukey's biweight estimate of the median; (vi) Peng's method [24], which performs a MA-data rotation step followed by LOWESS normalization.

NimbleGen uses Tukey's biweight scaling in-house. It consists of two steps: calculating the log-ratio between channels and subsequently correcting these by subtracting the robust Tukey's biweight estimate of the median. For this estimate, each data point is given a weight using a bi-square function. The weights assigned by this function are inversely correlated to the distance from the median, so outliers have a minimal effect on the estimate. The method developed by Peng et al. [24] makes strong assumptions regarding the shape of the MA-plot. In this approach, LOWESS normalization is preceded by a rotation step of the MA-data, which is meant to account for major dye trends. This method has been mostly applied in Drosophila [24,37,38].

Quantifying the effect of normalization on the two-component distribution

The separation between the enriched and un-enriched components present in the data of two-channel regulation microarrays should be conserved after applying normalization. To determine this conservation, the log-ratio distribution of negative control probes (which are a measure of non-specific annealing and background fluorescence) and the log-ratio distribution of gene promoter probes were assessed before and after normalization using ROC curves. For creating the ROC curves, the negative control probes represent the negative class of outcomes, while the gene promoter probes represent the positive class of outcomes. If there are any enriched probes, the gene promoter probe distribution should extend beyond the control probe distribution in the upper quantile and is expected to have a higher mean than the control probe distribution. If this separation is retained, the ROC curves are expected to have comparable AUC values before and after normalization, while if the separation is not retained, the ROC curves will have lower AUC values after normalization.

Genomic regions known *a priori* to be enriched were used as positive controls, verifying to what extent these regions are identified after using each of the six normalization approaches. To this end, 33 well established ER-a targets [25] were chosen as positive controls for dataset #1 [26]. Enrichment of these targets was calculated using ACME with default settings and a sliding window of 750 bp [11]. For dataset #3 enrichment of the HOXA group of developmental genes was determined, which are located in a cluster on chromosome 7 and are known to be silenced and moderately to highly methylated in most tissues [27]. Enrichment p-values were calculated with ACME using default settings and a sliding window of 1000 bp. For dataset #4 the same approach was used, focussing on the Dlk1-Gtl2 cluster on chromosome 12, a region that was identified as highly methylated in the original results [28]. The other datasets lacked suitable positive controls.

Data was imported and analyzed using Bioconductor [23] in the statistical programming language R, more specifically using the ACME package [11] for enrichment finding, the limma package [39] for data normalization and the Ringo package [40] for data import and handling.

Author contributions

MA and MJ carried out the analysis and interpretation of the microarray data and wrote the manuscript. CM, LE and CE helped drafting the manuscript and advised on the technical aspects and interpretation of the results. CE coordinated the project. All authors have read and approved the final manuscript.

Acknowledgements

Part of this work was carried out within the research program of the Netherlands Consortium for Systems Biology (NCSB), which is part of the Netherlands Genomics Initiative (NGI). Early stages of this work were supported by the European Nutrigenomics Organization (NuGO).

References

1. Zheng M, Barrera LO, Ren B, Wu YN: ChIP-chip: Data, Model, and Analysis. Biometrics 2007, 63(3):787-796.
2. Mohn F, Weber M, Schübeler D, Roloff T-C: Methylated DNA Immunoprecipitation (MeDIP). Methods Mol Biol 2009, 507:55-64.
3. Ordway JM, Bedell JA, Citek RW, Nunberg A, Garrido A, Kendall R, Stevens JR, Cao D, Doerge RW, Korshunova Y et al: Comprehensive DNA methylation profiling in a human cancer genome identifies novel epigenetic targets. Carcinogenesis 2006, 27(12):2409-2423.
4. Ballestar E, Paz MF, Valle L, Wei S, Fraga MF, Espada J, Cigudosa JC, Huang TH-M, Esteller M: Methyl-CpG binding proteins identify novel sites of epigenetic inactivation in human cancer. The EMBO journal 2003, 22(23):6335-6345.
5. Esteller M: Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet 2007, 8(4):286-298.
6. Movassagh M, Choy M-K, Goddard M, Bennett MR, Down TA, Foo RSY: Differential DNA methylation correlates with differential expression of angiogenic factors in human heart failure. PloS one 2010, 5(1):e8564.
7. Mariman ECM: Epigenetic manifestations in diet-related disorders. Journal of nutrigenetics and nutrigenomics 2008, 1(5):232-239.
8. Tsankova N, Renthal W, Kumar A, Nestler EJ: Epigenetic regulation in psychiatric disorders. Nature reviews Neuroscience 2007, 8(5):355-367.
9. Aryee MJ, Wu Z, Ladd-Acosta C, Herb B, Feinberg AP, Yegnasubramanian S, Irizarry RA: Accurate genome-scale percentage DNA methylation estimates from microarray data. Biostatistics (Oxford, England) 2011, 12(2): 197-210.
10. Johannes F, Wardenaar R, Colomé-Tatché M, Mousson F, de Graaf P, Mokry M, Guryev V, Timmers HTM, Cuppen E, Jansen RC: Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq. Bioinformatics (Oxford, England) 2010, 26(8):1000-1006.
11. Scacheri PC, Crawford GE, Davis S: Statistics for ChIP-chip and DNase hypersensitivity experiments on NimbleGen microarrays. Methods Enzymol 2006, 411:270-282.
12. Yang YH, Dudoit S, Luu P: Normalization for cDNA microarray data. Optical Technologies and Informatics 2001, 4266:141-152.
13. Song JS, Johnson WE, Zhu X, Zhang X, Li W, Manrai AK, Liu JS, Chen R, Liu XS: Model-based analysis of two-color microarrays (MA2C). Genome Biol 2007, 8:R178.
14. Lu R, Lee G-C, Shultz M, Dardick C, Jung K, Phetsom J, Jia Y, Rice RH, Goldberg Z, Schnable PS et al: Assessing probe-specific dye and slide biases in two-color microarray data. BMC Bioinformatics 2008, 9:314.

15. Wu Z, Irizarry RA, Gentleman R, Murillo FM, Spencer F: A Model-Based Background Adjustment for Oligonucleotide Expression Microarrays. Journal of the American Statistical Association 2004, 99:909-917.
16. Smyth GK, Speed T: Normalization of cDNA microarray data. Methods 2003, 31(4):265-273.
17. Kepler TB, Crosby L, Morgan KT: Normalization and analysis of DNA microarray data by self-consistency and local regression. Genome Biol 2002, 3(7):RESEARCH0037.
18. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielser HBr, Saxild H-H, Nielsen C, Brunak Sr, Knudsen S: A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome biology 2002, 3(9):research0048.
19. Bolstad BM, Irizarry RA, Astrand M, Speed TP: A comparison of normalization methods for high density oligonucleotide microarray data based on variance and bias. Bioinformatics (Oxford, England) 2003, 19(2):185-193.
20. Dudoit S, Yang YH, Callow MJ: Statistical methods for identifying differentially expressed genes in replicated cDNA microarray. Statistica Sinica 2002, 12:111-139.
21. Benoukraf T, Cauchy P, Fenouil R, Jeanniard A, Koch F, Jaeger Sb, Thieffry D, Imbert J, Andrau J-C, Spicuglia S, Ferrier P: CoCAS: a ChIP-on-chip analysis suite. Bioinformatics (Oxford, England) 2009, 25(7):954-955.
22. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M: Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics (Oxford, England) 2002, 18 Suppl 1:S96-104.
23. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J et al: Bioconductor: open software development for computational biology and bioinformatics. Genome biology 2004, 5(10):R80.
24. Peng S, Alekseyenko AA, Larschan E, Kuroda MI, Park PJ: Normalization and experimental design for ChIP-chip data. BMC bioinformatics 2007, 8:219.
25. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoute J, Brodsky AS, Keeton EK, Fertuck KC, Hall GF et al: Genome-wide analysis of estrogen receptor binding sites. Nat Genet 2006, 38(11):1289-1297.
26. Romano A, Adriaens M, Kuenen S, Delvoux B, Dunselman G, Evelo C, Groothuis P: Identification of novel ER-alpha target genes in breast cancer cells: gene- and cell-selective co-regulator recruitment at target promoters determines the response to 17beta-estradiol and tamoxifen. Mol Cell Endocrinol 2009, 314(1):90-100.
27. Hayashi H, Nagae G, Tsutsumi S, Kaneshiro K, Kozaki T, Kaneda A, Sugisaki H, Aburatani H: High-resolution mapping of DNA methylation in human genome using oligonucleotide tiling array. Human genetics 2007, 120(5):701-711.
28. Wu H, Tao J, Chen PJ, Shahab A, Ge W, Hart RP, Ruan X, Ruan Y, Sun YE: Genome-wide analysis reveals methyl-CpG-binding protein 2-dependent regulation of microRNAs in a mouse

model of Rett syndrome. Proceedings of the National Academy of Sciences of the United States of America 2010, 107(42):18161-18166.

29. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic acids research 2002, 30(4):e15.
30. Schadt EE, Li C, Ellis B, Wong WH: Feature extraction and normalization algorithms for high-density oligonucleotide gene expression microarray data. Journal of cellular biochemistry Supplement 2001, Suppl 37:120-125.
31. Wilson AS, Power BE, Molloy PL: DNA hypomethylation and human diseases. Biochim Biophys Acta 2007, 1775(1):138-162.
32. Zhao Y, Li M-C, Simon R: An adaptive method for cDNA microarray normalization. BMC bioinformatics 2005, 6:28.
33. Pelizzola M, Koga Y, Urban AE, Krauthammer M, Weissman S, Halaban R, Molinaro AM: MEDME: an experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment. Genome research 2008, 18(10):1652-1659.
34. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Gräf S, Johnson N, Herrero J, Tomazou EM: A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat Biotechnol 2008, 26(7):779-785.
35. McKay JA, Adriaens ME, Ford D, Relton CL, Evelo CTA, Mathers JC: Bioinformatic interrogation of expression microarray data to identify nutritionally regulated genes potentially modulated by DNA methylation. Genes Nutr 2008, 3(3-4):167-171.
36. Kauffmann A, Gentleman R, Huber W: arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. Bioinformatics (Oxford, England) 2009, 25(3):415-416.
37. Sural TH, Peng S, Li B, Workman JL, Park PJ, Kuroda MI: The MSL3 chromodomain directs a key targeting step for dosage compensation of the Drosophila melanogaster X chromosome. Nature structural & molecular biology 2008, 15(12):1318-1325.
38. Gelbart ME, Larschan E, Peng S, Park PJ, Kuroda MI: Drosophila MSL complex globally acetylates H4K16 on the male X chromosome for dosage compensation. Nature structural & molecular biology 2009, 16(8):825-832.
39. Smyth GK, Speed TP: Normalization of cDNA microarray data. Methods 2003, 31:265-273.
40. Toedling J, Sklyar O, Huber W: Ringo - an R/Bioconductor package for analyzing ChIP-chip readouts. BMC Bioinformatics 2007, 8(1):221.

Chapter 4

Capturing ChIP-seq profiles of H3K27me3 in dynamic biological systems

Michiel Adriaens[1]*, Peggy Prickaerts[2]*, Michelle Chan-Seng-Yue[3], Timothy Beck[3], Bradly G Wouters[3,4,5,6,#], Jan Willem Voncken[2,#], Chris Evelo[1,#]

** Equal contribution.*

*** To whom correspondence may be addressed*

[1] Department of Bioinformatics – BiGCaT, Maastricht University, Maastricht, The Netherlands

[2] Department of Molecular Genetics, Maastricht University, Maastricht, The Netherlands

[3] Ontario Institute for Cancer Research, Toronto, ON, Canada

[4] Ontario Cancer Institute and Campbell Family Institute for Cancer Research, Princess Margaret Hospital, University Health Network, Toronto, Canada

[5] Maastricht Radiation Oncology (MaastRO) Lab, Maastricht University, Maastricht, The Netherlands

[6] Departments of Radiation Oncology and Medical Biophysics, University of Toronto, Toronto, ON, Canada

Keywords: *ChIP-seq, normalization, bioinformatics, epigenetics, H3K27me3, blanketing.*

Publication: *Submitted*

Abstract

Chromatin immunoprecipitation combined with high-throughput sequencing (ChIP-seq) has become the state-of-the-art technology for genome-wide mapping of epigenetic modifications. ChIP-seq has a much higher resolution than microarray based applications; this impacts on the requirements for analysis tools. Here we focus on applying ChIP-seq to study enrichment of tri-methylation of the lysine 27 residue on histone H3 (H3K27me3) in dynamic biological systems, where the majority of epigenetic markings are different between samples. A number of specific challenges exist with regards to studying histone modifications under dynamic conditions. Firstly, ChIP-seq enrichment-finding algorithms are optimized for identifying sharp peaks, whereas H3K27me3 marks tend to spread over a large number of nucleosomes (blanketing) and cover extended genomic regions. Thus, existing algorithms are unable to reliably identify H3K27me3 enriched regions. Secondly, ChIP-seq data requires normalization to enable quantitative comparison. In dynamic biological systems, both the total number of marks and their location is variable. A suitable normalization approach in this situation involves data-scaling of individual samples to genomic regions with relatively stable H3K27me3 marking. Definition of such stable regions depends heavily on the biological system. With these difficulties in mind, we have developed a standardized protocol for the processing of H3K27me3 ChIP-seq data. The protocol enables robust detection of H3K27me3 blanketing and allows for quantitative data comparison. As such, our protocol complements previous efforts to create a fully standardized analysis-pipeline for H3K27me3 enriched ChIP-seq data.

Introduction

High-throughput sequencing (HTS) has enabled the study of biological systems on a greater scale and higher genomic resolution than ever before [1] and is prospected to soon replace microarrays as the platform of choice. Although the technology is constantly evolving, analysis approaches are lagging behind [2]. Most approaches are derivatives of approaches developed for the corresponding application on microarrays [3], which are characterized by lower genomic resolution and scale.

Epigenomics is the genome-wide study of epigenetic modifications, such as DNA methylation and modification of N-terminal histone tails, which are heritable but nonetheless variable marks that influence gene expression and phenotypic plasticity. Well known examples of histone modifications are tri-methylation of lysine 4 of histone H3 (H3K4me3), which is associated with open chromatin and permissive for gene transcription, and tri-methylation of lysine 27 of histone H3 (H3K27me3), which is associated with closed chromatin and gene silencing. Chromatin immunoprecipitation (ChIP) combined with HTS technology (ChIP-seq) is a powerful approach to create genome-wide maps of such histone modifications. Here we focus specifically on H3K27me3 ChIP-seq data analysis in the context of biological systems where many epigenetic changes occur. Cancer cells exposed to fluctuating levels of oxygenation exemplify a model for such dynamic effects occurring in response to hypoxia in solid tumors.

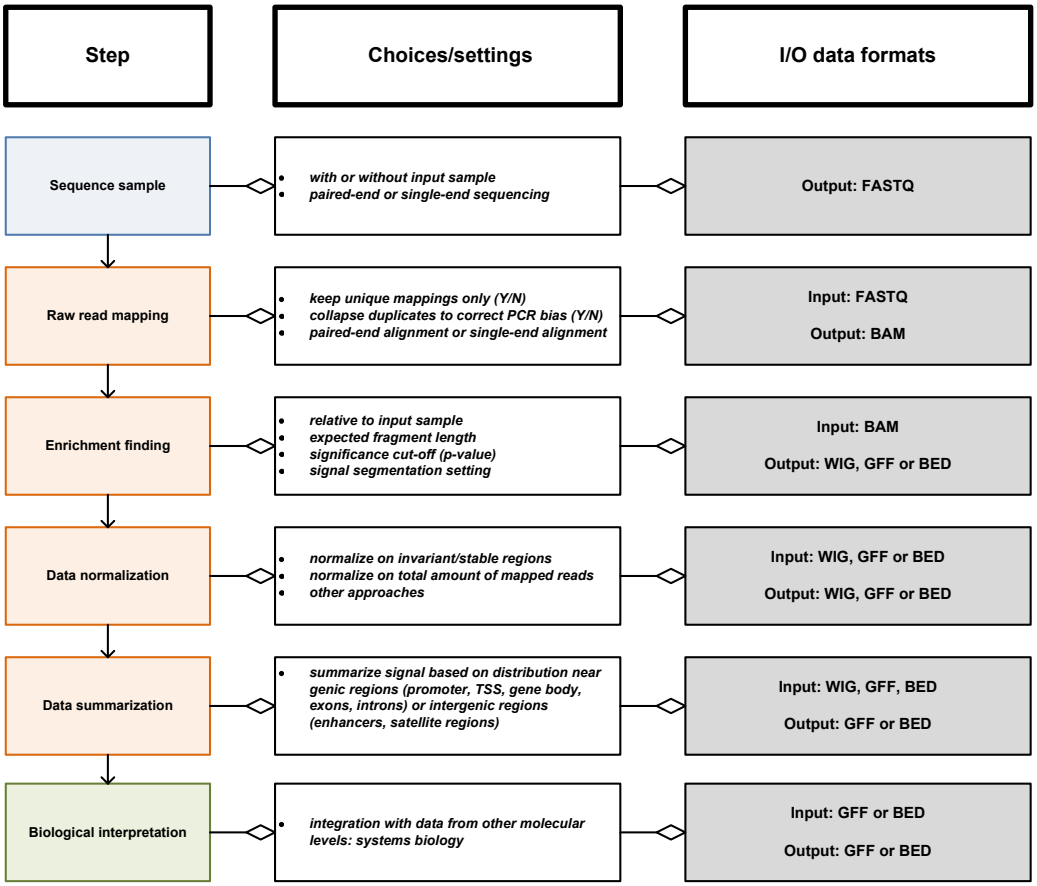


Figure 1: *Flow diagram of steps involved in the bioinformatics analyses of ChIP-seq data, from raw reads to biological interpretation.*

The sequential steps involved in ChIP-seq analysis are summarized in **figure 1**. Each step is paralleled by important decisions and specific data formats. This flow scheme concludes with interpretation of the data in a biologically relevant context.

Step 1: Sequencing

The steps before enrichment finding are well established and are independent of the biological system studied. Yet a number of important considerations apply to the design of the sequencing experiment. The first is whether or not to include an input sample in parallel as a reference for analysis of ChIP-seq data. Although including input samples increases overall costs and reduces the number of experimental samples in one run, an input sample is absolutely essential for proper correction of background anomalies, such as amplified regions and variability in shearing of DNA [4]. Also the choice of sequencing

technology, *i.e.* paired-end or single-end sequencing, requires consideration: paired-end sequencing yields paired reads corresponding to the start and the end of a fragment and thus allows for more robust alignment. For ChIP-seq experiments, paired-end technology is clearly the preferred method, provided it is available [5].

Step 2: Mapping reads

The next step in the process is the mapping of the raw sequence reads to a reference genome build, most commonly performed using Needleman-Wunsch or BLAST algorithms [6]. Some are designed for rapid analysis, such as Eland, while others aim for sensitivity, such as Novoalign [7]. Important decisions to make at this step include exclusion of non-unique mappings and collapsing duplicate reads caused by PCR bias. In most cases, collapsing and proceeding with unique mappings only leads to more reliable, unbiased data [7]. Depending on the technology applied, raw reads consist of paired-end reads or single-end reads. As paired-end sequencing provides two connected DNA end-tags, it enables reliable identification of enriched regions. In addition, paired-end sequencing is more powerful in identifying enrichment in repeat regions, such as satellite DNA regions near centromeres. Single-end mappings from such regions will often be discarded from the analysis, because they do not map uniquely to one defined region. In case of paired-end reads, two close mappings, with ideally only one out of two reads mapping into a repeat region, prevent loss of such repeat regions. Since several histone modifications are associated with such regions and as such are of biological interest [8], paired-end mapping is always the algorithm of choice when studying histone modifications [5].

Step 3: Identifying regions of enrichment

Several tools are available for enrichment-finding in ChIP-seq data, of which FindPeaks [9], PeakSeq [10], USeq [11] and MACS [12] are the most popular. Although these tools produce comparable results, FindPeaks is most sensitive in distinguishing enrichment [13]. Enrichment-finding tools work by the assumption that regions enriched for a histone modification of interest will yield a higher number of reads representing the regions in ChIP samples relative to un-enriched regions. Hence, significantly enriched regions can be identified [14] against a null distribution, ideally in the form of a sequenced input sample. Additional modulations to support enrichment-finding include modification of the expected fragment-length (post-sonication), resetting the significance cut-off (usually in the form of a percentile of the data above which a signal is considered enriched; e.g. 95%), and varying the signal-segmentation setting (determines when a multi-modal signal should be split into separate peaks). Most of these distinct enrichment-finding approaches have been developed for lower resolution cistromics applications on microarray platforms. Although they work well for many histone modifications deposited over a small number of nucleosomes such as H3K4me3, such approaches are not optimized for detection of H3K27me3 enrichment. A potential underlying cause is that H3K27me3 tends to spread over more nucleosomes to cover an entire locus, which is known as “blanketing” [15]. This results in data comprising

spread-out enrichment signals covering large genomic regions [15,16,17] instead of sharp peaks, which hampers their detection by existing peak-finding algorithms.

Step 4: Data normalization

To compare samples quantitatively, ChIP-seq data requires normalization, because signal intensity depends on sequencing depth and mapping efficiency. Several approaches have been developed, of which scaling based on the total number of aligned reads is most common [18,19]. This method starts with the assumption that there is only a small number of prospected differences between conditions, which does not hold for dynamic biological systems; for such systems normalizing on regions with stable enrichment seems the only valid approach. However, it is difficult to define *a priori* genomic regions as candidates for such stable enrichment, as this relies heavily on the biological system as well as the studied histone modification [18].

Step 5 & 6: Data summarization and interpretation.

Although usually complex and mostly dependent on the biological system of choice and research question(s) at hand, the biological interpretation of the data is identical to comparable applications in microarrays [20]: integration of transcriptomics data is the minimal requirement for a meaningful analysis. Since experiments and research questions are extremely diverse, a flexible analysis tool is needed. Perhaps not surprising, several open source tools are available to enable robust analysis and biological interpretation of ChIP-seq data [21,22,23]. However, the sheer amount of sequencing information obstructs straight-forward data interpretation. Hence, most researchers choose to focus on genes and/or regulatory regions before committing to the biological interpretation phase. This requires summarization of the enrichment in genic and regulatory regions of interest. For H3K27me3 data, such regulatory relevant regions have been defined successfully before [15].

Based on the considerations above, we have developed a protocol to handle H3K27me3 ChIP-seq data focused on (i) standardization of the enrichment finding, (ii) data normalization and (iii) data summarization steps, while building on the already standardized stages before enrichment finding.

Procedure

(i) Finding regions of enrichment

H3K27me3 is known to "blanket" over large genomic regions. In order to define such broad regions of enrichment, while still retaining all the advantages of using an input DNA sample as a reference, FindPeaks with adapted settings is used. The primary setting that needs to be changed in the FindPeaks algorithm is the expected sequence distribution setting. By increasing the expected fragment size (post-sonication), the algorithm successfully classifies broad regions of enrichment as peaks (**figure 2**).

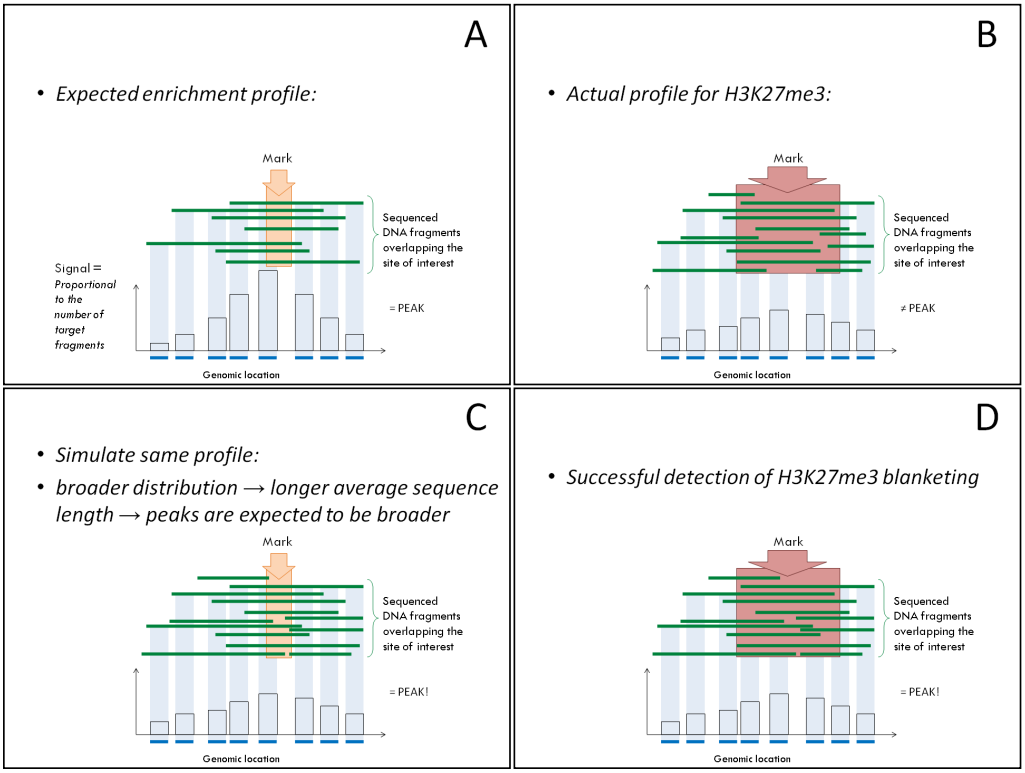


Figure 2: *Illustration of the effect of increasing the expected fragment length-distribution to enable more robust detection of H3K27me3 enriched regions.*

Additional settings are tied to the normalization approach: for this method to work, the signal needs to be split in as many individual peaks as possible, which is controlled by the signal-segmentation setting. Furthermore, the approach requires retention of as much of the signal as possible, which is controlled by the significance cut-off setting. Although this will inevitably increase the size of the resulting data files, as

the identified peaks will be broader, the signal sampling frequency can safely be decreased as a counter measure.

Using the mapped input reads as control sample with flag "-control" and the immunoprecipitated sample as the experimental sample with the flag "-input", we use the following "broad-peak" settings:

- *-aligner sam* *(aligner type)*
- *-subpeaks 0.5* *(signal segmentation)*
- *-control_type 0* *(control type: input DNA)*
- *-alpha 0.2* *(significance setting)*
- *-wig_step_size 10* *(signal sampling frequency)*
- *-trim 0.2* *(peak trimming)*
- *-dist_type 1 285 600 100* *(expected sequence distribution)*

As an example, using FindPeaks with these settings would lead to the following terminal command*:*

```
java -jar  /test/local/analysis/src/findPeaks/fp4.0.16/fp4/FindPeaks.jar
       -aligner sam
       -hist_size 100
       -input ../5.merge.filtered4_256.collapsed.q30.bam
       -name 5.merge.filtered4_256.collapsed.q30.bam.out
       -control ../../../1/novoalign/1.merge.filtered4_256.collapsed.q30.bam
       -output .
       -subpeaks 0.5
       -control_type 0
       -alpha 0.2
       -wig_step_size 10
       -trim 0.2
       -dist_type 1 285 600 100
```

(ii) Data normalization

Normalization follows the identification of H3K27me3-enriched signals that are significantly above background level. The normalization strategy is based on identifying regions with stable H3K27me3-enrichment between two or more samples. The cumulative area under the curve (AUC) for all peaks in all these regions is determined for each condition, and based on those values, scaling factors are calculated relative to the smallest value among the samples (**figure 3**).

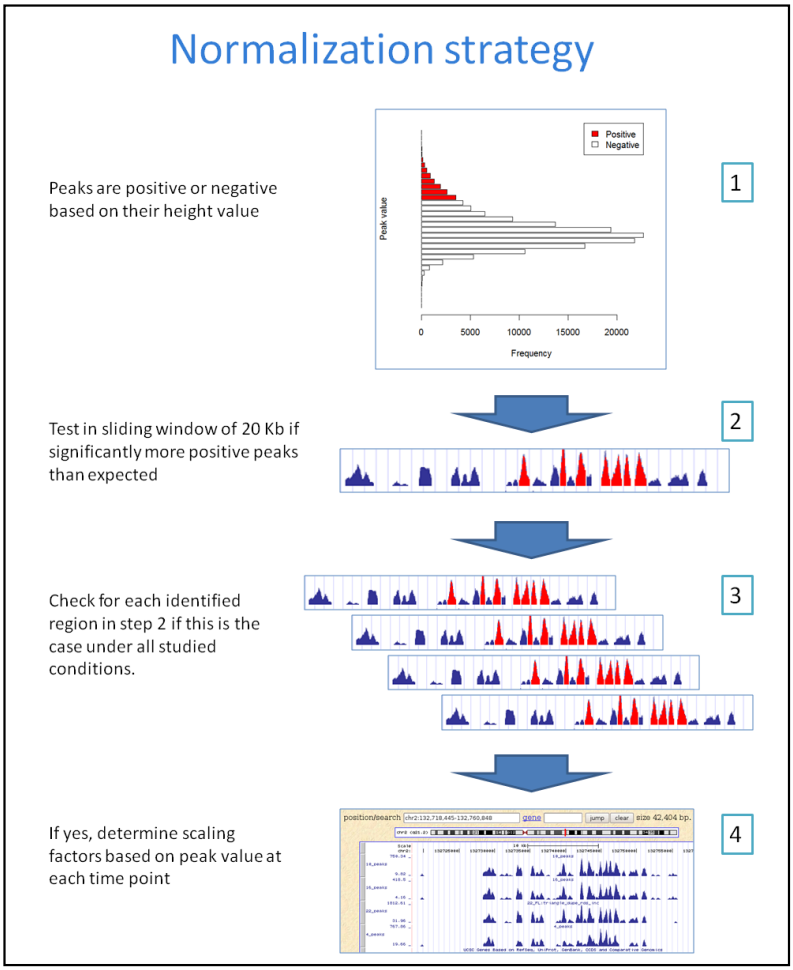


Figure 3: *A schematic overview showing the steps involved in the normalization strategy based on finding regions with stable H3K27me3 enrichment.*

The most straightforward approach is to use the ACME package [14] in R, using the following approach:

- Import the files containing peak data in R, including for each peak at least a genomic location and a height or AUC-value (GFF or BED format).
- Then for all samples/conditions:
    - Create an ACME data object as described in the ACME documentation [14]
    - Run ACME using a window size of at least 20000 bp and a cut-off of 0.90 corresponding to the peak height value of the 90th percentile. Peaks with a height above this value will be designated positive, those below this value will be designated negative (**figure 3**). The

window size of 20000 bp is chosen because it is much larger than the observed average peak width when using the settings specified in step (i), while being small enough to lead to a reasonable specificity.

    - Upon completion of the run, proceed only with regions with a chi-square test p-value smaller than 1E-5 in all samples/conditions. This value was shown empirically to yield good results.

- Add up all signals detected within the stable regions for each sample individually.
- Divide the individual sums by the smallest sum in the batch and use the resulting scaling factors to scale the peaks for each sample individually.

(iii) Summarization at gene level

To ease the interpretation of high-scale datasets such as obtained in ChIP-seq analysis, an excellent approach is to summarize the data based on relative location with respect to genes and gene regulatory regions. For H3K27me3, such regions have been defined successfully before, with clear associations to gene transcription regulation [15]:

- Promoter region: 3000 bp upstream of TSS to 100 bp upstream of TSS
- TSS region: 100 bp upstream of TSS to 1000 bp downstream of TSS
- Gene body: 1000 bp downstream of TSS to end of the last exon

Enrichment for H3K27me3 that covers the entire gene body is associated with gene silencing. Enrichment within the TSS region are candidates for so-called bivalency, *i.e.* nucleosomal regions marked with both gene-silencing-associated H3K27me3 and gene-activity-associated H3K4me3. Enrichment in the promoter region can be associated with gene silencing as well as active gene transcription, depending on the nearby presence of other histone modifications and the cellular context [15].
Other regions of interest (ROI) may be enhancers, which are known to be susceptible to various histone modifications [24]. A potential obstacle with these ROIs is that clear definitions of what constitutes a functional enhancer is not always available, as these are functionally dependent on cellular context.

Anticipated results

We have tested the above protocol extensively on our own and published H3K27me3 ChIP-seq datasets. The enrichment-finding approach and the normalization approach are inseparably connected. The broad-peak settings for FindPeaks produce more returned signal, which is separated into more individual peaks (**figure 4**).

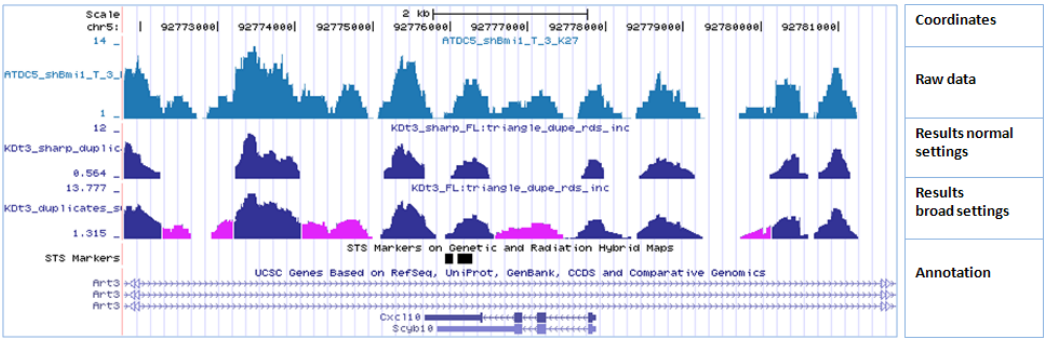


Figure 4: *Results of the broad-peak settings for FindPeaks compared to the default settings.*

This is required for the normalization-approach to work properly, because it needs to integrate as much peak information as possible to reliably identify stable regions of enrichment within the genome. For the dataset used to develop the protocol, most of these stable regions were found near centromeres and in intergenic regions. It has been observed previously that centromeres are heavily enriched for H3K27me3 [8,25,26,27]. These regions are also known to contain a large number of repeats. Therefore, repeat region reads will inevitably end-up being mapped to these regions, regardless of their origin. Hence, the reason for finding high enrichment within these regions may reflect a technical artifact rather than biology. It is therefore essential to prevent accumulation of such reads leading to spurious enrichment in these regions, by keeping only uniquely mapped reads for normalization analysis.

We compared the results for our dataset and approach to published results [28]. In **figure 5**, condensed genome tracks are plotted for chromosome 1 and 19, showing very similar profiles, although our approach retains more signal.

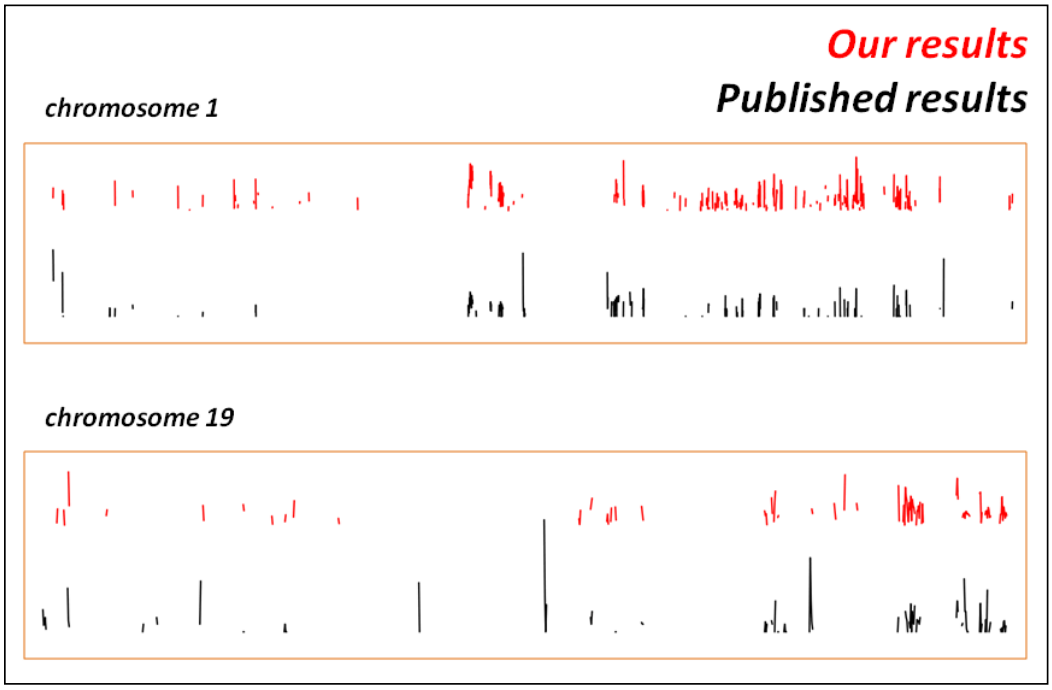


Figure 5: *Comparison of published results with our results: a similar "fingerprint" is found, but more signal is retained using our H3K27me3 enrichment protocol.*

Discussion

We here report on a straight forward, robust method to delineate non-uniform genomic H3K27me3-distributions. Our approach integrates a normalization method based on identifying regions with stable H3K27me3-enrichment between samples and a cumulative signal-retention strategy, which identifies broadly defined epigenetic marks, which would be lost with standard peak-find algorithms.

With respect to definition of H3K27me3-enrichment in ChIP-seq data multiple approaches are in use, which mostly build on limiting the enrichment-finding step to predefined regions and counting the number of reads in these regions. Some methods apply this to genic regions only [15,29], some to intergenic regions only [27] and others to both [30]. By limiting the enrichment-finding to regions with predefined start and end points, however, the enrichment analysis requires redefinition each time when novel regions of interest are probed, such as enhancers which are often located in the predefined intergenic regions. Moreover, when neither input sample data nor an estimated null-distribution is considered in the enrichment–finding strategy, which is the case for some published approaches [29,30], this can result in spurious background enrichment, which limits the reliability of quantitative comparisons, regardless of normalization. Another downside of many of these approaches is that definition of enrichment is data driven, resulting in analytical parameter use that may not be applicable to other studies.

Genome-wide enrichment-finding methodology not limited by pre-defined static regions, include, for instance, a hidden Markov-model approach. Although some methods compare H3K27me3-enrichment between samples in a genome-wide fashion [31,32] and/or apply a genome-wide sliding window approach (*i.e.* summarizing the counts in each window; [33,34]), none of these approaches integrate input sample data or an estimated null-distribution to correct for background anomalies. Several studies report the use of standard enrichment-finding algorithms, like MACS [35], in combination with input sample data [36,37,38]. Using such algorithms without adapted settings, however, will inevitably result in significant loss of H3K27me3 signal.

With regard to data normalization, the most often reported approach is scaling by the total number of unique reads of each sample to the sample with the lowest number of reads [15,30,32]. A related approach is the dissection of nucleosome-filled genomic portions (or a subset thereof) into non-overlapping windows, followed by dividing the count-number in each window by the average number of counts for all windows [27,29]. Note that both normalization methods start from the assumption that the number of (any given) histone modifications does not vary between conditions. Whereas this may be satisfactory for specific experimental settings, it does not apply to dynamic biological systems. Strikingly, many published reports have omitted any form of data normalization as part of the data summarizing flow [33,34,36,37,38].

In conclusion, to our knowledge, the herein described approach is the only generic protocol that enables genome-wide, quantitative data comparison for H3K27me3 ChIP-seq data, while still retaining the robustness of using input sample data for enrichment-finding. Other histone modifications exist with similar distribution-profiles as H3K27me3, an example of which is the H3K36me3 modification; H3K36me3 is associated with transcription-elongation and typically covers the most if not all of the gene body. Hence, while our method is designed and optimized for studying H3K27me3-histone modifications, the protocol is readily adaptable for definition of other broad-type epigenetic enrichment profiles.

R scripts are available from the authors upon request.

References

1. Mardis ER: ChIP-seq: welcome to the new frontier. Nature methods 2007, 4(8):613-614.
2. Park PJ: ChIP-seq: advantages and challenges of a maturing technology. Nature reviews Genetics 2009, 10(10):669-680.
3. Ho JWK, Bishop E, Karchenko PV, Nègre N, White KP, Park PJ: ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. BMC genomics 2011, 12:134.
4. Kharchenko PV, Tolstorukov MY, Park PJ: Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nature biotechnology 2008, 26(12):1351-1359.
5. Fullwood MJ, Wei C-L, Liu ET, Ruan Y: Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome research 2009, 19(4):521-532.

6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. Journal of molecular biology 1990, 215(3):403-410.
7. Ruffalo M, LaFramboise T, Koyutürk M: Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics (Oxford, England) 2011, 27(20):2790-2796.
8. Day DS, Luquette LJ, Park PJ, Kharchenko PV: Estimating enrichment of repetitive elements from high-throughput sequence data. Genome biology 2011, 11(6):R69.
9. Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones SJM: FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics (Oxford, England) 2008, 24(15):1729-1730.
10. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB: PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nature biotechnology 2009, 27(1):66-75.
11. Nix DA, Courdy SJ, Boucher KM: Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. BMC bioinformatics 2008, 9:523.
12. Zhang Y, Liu T, Meyer CA, Eeckhoute Jrm, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al: Model-based analysis of ChIP-Seq (MACS). Genome biology 2008, 9(9):R137.
13. Malone BM, Tan F, Bridges SM, Peng Z: Comparison of four ChIP-Seq analytical algorithms using rice endosperm H3K27 trimethylation profiling data. PloS one 2011, 6(9):e25260.
14. Scacheri PC, Crawford GE, Davis S: Statistics for ChIP-chip and DNase hypersensitivity experiments on NimbleGen arrays. Methods Enzymol 2006, 411:270-282.
15. Young MD, Willson TA, Wakefield MJ, Trounson E, Hilton DJ, Blewitt ME, Oshlack A, Majewski IJ: ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. Nucleic acids research 2011, 39(17):7415-7427.
16. Bracken AP, Dietrich N, Pasini D, Hansen KH, Helin K: Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. Genes & development 2006, 20(9):1123-1136.
17. Pietersen AM, van Lohuizen M: Stem cell regulation by polycomb repressors: postponing commitment. Current opinion in cell biology 2008, 20(2):201-207.
18. Huang W, Umbach DM, Vincent Jordan N, Abell AN, Johnson GL, Li L: Efficiently identifying genome-wide changes with next-generation sequencing data. Nucleic acids research 2011, 39(19):e130.
19. Cheung M-S, Down TA, Latorre I, Ahringer J: Systematic bias in high-throughput sequencing data and its correction by BEADS. Nucleic acids research 2011, 39(15):e103.
20. Adriaens ME, Romano A, Eijssen LMT, McKay JA, Mathers JC, Evelo CTA. enrichR: a Bioconductor toolbox for integrative analysis of ChIP-on-chip and DNA methylation microarray data. *[submitted]*
21. Taslim C, Huang T, Lin S: DIME: R-package for identifying differential ChIP-seq based on an ensemble of mixture models. Bioinformatics (Oxford, England) 2011, 27(11):1569-1570.

22. Muiño JM, Kaufmann K, van Ham RC, Angenent GC, Krajewski P: ChIP-seq Analysis in R (CSAR): An R package for the statistical detection of protein-bound genomic regions. Plant methods 2011, 7:11.
23. Mercier E, Droit A, Li L, Robertson G, Zhang X, Gottardo R: An integrated pipeline for the genome-wide analysis of transcription factor binding sites from ChIP-Seq. PloS one 2011, 6(2):e16432.
24. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW et al: Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 2009, 459(7243):108-112.
25. Saurin AJ, Shiels C, Williamson J, Satijn DP, Otte AP, Sheer D, Freemont PS: The human polycomb group complex associates with pericentromeric heterochromatin to form a novel nuclear domain. The Journal of cell biology 1998, 142(4):887-898.
26. Voncken JW, Schweizer D, Aagaard L, Sattler L, Jantsch MF, van Lohuizen M: Chromatin-association of the Polycomb group protein BMI1 is cell cycle-regulated and correlates with its phosphorylation status. Journal of cell science 1999, 112 (Pt 24):4627-4639.
27. Rosenfeld JA, Wang Z, Schones DE, Zhao K, DeSalle R, Zhang MQ: Determination of enriched histone modifications in non-genic portions of the human genome. BMC genomics 2009, 10:143.
28. Joseph R, Orlov YL, Huss M, Sun W, Kong SL, Ukil L, Pan YF, Li G, Lim M, Thomsen JS et al: Integrative model of genomic factors for determining binding site selection by estrogen receptor-α. Molecular systems biology 2010, 6:456.
29. Chopra VS, Hendrix DA, Core LJ, Tsui C, Lis JT, Levine M: The polycomb group mutant esc leads to augmented levels of paused Pol II in the Drosophila embryo. Molecular cell 2011, 42(6):837-844.
30. Marks H, Chow JC, Denissov S, Françoijs K-J, Brockdorff N, Heard E, Stunnenberg HG: High-resolution analysis of epigenetic changes associated with X inactivation. Genome research 2009, 19(8):1361-1373.
31. Xu H, Sung W-K: Identifying Differential Histone Modification Sites from ChIP-seq Data. Methods in molecular biology (Clifton, NJ) 2012, 802:293-303.
32. Xu H, Wei C-L, Lin F, Sung W-K: An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. Bioinformatics (Oxford, England) 2008, 24(20):2344-2349.
33. Pauler FM, Sloane MA, Huang R, Regha K, Koerner MV, Tamir I, Sommer A, Aszodi A, Jenuwein T, Barlow DP: H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. Genome research 2009, 19(2):221-233.
34. Akkers RC, van Heeringen SJ, Jacobi UG, Janssen-Megens EM, Françoijs K-J, Stunnenberg HG, Veenstra GJC: A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in Xenopus embryos. Developmental cell 2009, 17(3):425-434.
35. Zhang Y, Liu T, Meyer CA, Eeckhoute Jrm, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al: Model-based analysis of ChIP-Seq (MACS). Genome biology 2008, 9(9):R137.

36. Kim SW, Yoon S-J, Chuong E, Oyolu C, Wills AE, Gupta R, Baker J: Chromatin and transcriptional signatures for Nodal signaling during endoderm formation in hESCs. Developmental biology 2011, 357(2):492-504.
37. Li H, Bitler BG, Vathipadiekal V, Maradeo ME, Slifker M, Creasy CL, Tummino PJ, Cairns P, Birrer MJ, Zhang R: ALDH1A1 is a novel EZH2 target gene in epithelial ovarian cancer identified by genome-wide approaches. Cancer prevention research (Philadelphia, Pa) 2011.
38. Suzuki H, Takatsuka S, Akashi H, Yamamoto E, Nojima M, Maruyama R, Kai M, Yamano H-O, Sasaki Y, Tokino T et al: Genome-wide profiling of chromatin signatures reveals epigenetic regulation of MicroRNA genes in colorectal cancer. Cancer research 2011, 71(17):5646-5658.

Chapter 5

The public road to high quality curated biological pathways

Michiel E. Adriaens[1], Magali Jaillard[1], Andra Waagmeester[1], Susan L.M. Coort[1], Alex R. Pico[2] and Chris T.A. Evelo[1]

[1]*Department of Bioinformatics - BiGCaT, Maastricht University, Maastricht, The Netherlands*
[2]*Gladstone Institute of Cardiovascular Disease, San Francisco, California, USA*

***Keywords**: Biological pathways, pathway analysis, biological databases, pathway curation.*

Publication: Adriaens, M.E., M. Jaillard, A. Waagmeester, S.L.M. Coort, A.R. Pico, C.T.A. Evelo (2008). "The public road to high-quality curated biological pathways." Drug Discov Today 13(19-20): 856-62.

Abstract

Biological pathways are abstract and functional visual representations of existing biological knowledge. By mapping high-throughput data on these representations, changes and patterns in biological systems on the genetic, metabolic and protein level are instantly assessable. Many public domain repositories exist for storing biological pathways, each applying its own conventions and storage format. A pathway based content review of these repositories reveals that none of them are comprehensive.

To address this issue, we apply a general workflow to create curated biological pathways, in which we combine three content sources: public domain databases, literature and experts. In this workflow all content of a particular biological pathway is manually retrieved from biological pathway databases and literature, after which this content is compared, combined and subsequently curated by experts. From the curated content new biological pathways can be created for a pathway analysis tool of choice and distributed amongst its user base. We applied this procedure to successfully construct high quality curated biological pathways involved in human fatty acid metabolism.

Introduction

In recent years biological pathway analysis has become common in biochemical research. There is a plethora of pathway analysis tools available. In general such tools map multidimensional experimental data to biological pathways, which are abstract and functional visual representations of existing biological knowledge [1]. Pathways help to understand changes and patterns in biological systems of various types of organisms on the genetic, protein and metabolic level.

A pathway may encompass one or several types of biological processes [2], the key ones being regulatory processes, metabolic processes, protein-protein interactions and signaling processes. A regulatory process may for instance involve transcription factors and the genes whose expression they activate or inhibit, an example of which is the regulation of genes involved in fatty acid metabolism by peroxisome proliferator-activated receptor alpha (PPAR-α). A metabolic process describes flows of physiological reactions, involving subtracts, products and commonly a catalyst, such as the series of reactions describing fatty acid β-oxidation. An example of a protein-protein interaction is the binding of a ligand to a receptor, which in turn may activate a signaling process like the mitogen activated protein kinase (MAPK) cascade. A pathway always shows direction, but it can contain more knowledge than a simple network, such as information on the sub-cellular localization of components, regulatory mechanisms and connections to other pathways.

Over the last decade several online databases were created to store biological pathways [3]. Each database has its own conventions, level of interactivity and storage format, but in the end all of them store information covering biological pathways, from the low-level processes of metabolism to high-level processes like regulation. Some databases store only a static picture of a biological pathway, such as the BioCarta database (http://www.biocarta.com/), while others, such as Reactome [4] and KEGG [5], store

extensive annotation for each of the elements in a pathway by using a custom XML format, in addition to a graphical layout. Another growing repository is Science Signaling (http://stke.sciencemag.org/), developed by American Association for the Advancement of Science (AAAS), the publisher of Science Magazine. It includes more than 60 curated signaling pathways and additionally provides short lists of key references and evaluations of the strength of existing evidence for associations within the database. WikiPathways (http://www.wikipathways.org/) applies a similar concept. It is a public Wiki platform dedicated to the creation, storing and curation of biological pathways by and for the scientific community. WikiPathways contains copies of all GenMAPP pathways plus additional pathways in GPML (GenMAPP Pathway Markup Language) format.
GeneGO Metacore [6], GenMAPP [7] and Metacyc (http://www.metacyc.org/) also offer data visualization and statistical analysis tools to analyze experimental data on a pathway level. This brings us to the most important application of pathways: pathway analysis.

Pathway analysis tools

Pathway analysis of gene or protein expression data applies genomic information to couple the expression data to known biological pathways. Usage of extensive collections of such pathways allows a quickly assessable overview of expression results in relation to biological mechanisms, facilitating the understanding of gene, protein and metabolite interactions at higher physiological levels. Cavalieri and De Filippo [8] reviewed tools that automatically display functional genomics results on biological pathways and tools that test for statistical significance of enrichment of genes belonging to a biological pathway. Among these tools are several commercial applications, such as GeneGO Metacore, Rosetta Resolver and Acuity, which are commonly used for high-throughput data analysis. Open-source programs such as GenMAPP with MAPPFinder [9], Cytoscape [10] and DAVID (http://david.abcc.ncifcrf.gov/) also offer the possibility of interactively visualizing expression datasets on biological pathways. When large amounts of expression data need to be analyzed on a large collection of pathways, a need for automation arises. Several statistical methods were developed to assess the significance of changes in gene expression in a pathway or a collection of pathways. Pathway analysis programs such as Pathway Miner [11], Eu.gene Analyzer [8], MetaCyc (http://www.metacyc.org/) and GenMAPP's MAPPFinder each have their own statistical approaches**.** An overview of some popular pathway editing and analysis tools is found in **table 1.**

GenMAPP is a popular freely available biological pathway analysis tool developed at the Conklin Lab at the J. David Gladstone Institutes of the University of California. Several gene properties can be displayed on pathways simultaneously by creating a lookup table, linking colours and descriptions to user specified criteria that for instance indicate changes in the gene expression level. GenMAPP's built-in statistical tool MAPPFinder enables a pathway-based enrichment analysis. MAPPFinder calculates a statistical p-value

for each pathway entity using the hypergeometrical distribution, after which pathways are ranked by significance.

	Pathway editing	Pathway analysis	Licence	Textmining[a]	Pathway databases
Biocarta	Yes	No	Free	No	Proprietary
EU.gene	Yes	Yes	Free	No	Proprietary, WikiPathways[b]
GenMAPP	Yes	Yes	Free	No	Proprietary, KEGG, WikiPathways[b]
Genomatix	No	Yes	Commercial	Yes	Proprietary
Ingenuity	Yes	Yes	Commercial	Yes	Proprietary, KEGG
MetaCore	Yes	Yes	Commercial	Yes	Proprietary
Pathway Studio	Yes	Yes	Commercial	Yes	ResNet Mammalian Database, ResNet Plant, ResNet Targeted Databases, KEGG, Science Signaling, Prolexys HyNet yeast two-hybrid database
Reactome	Yes	No	Free	No	Proprietary
WikiPathways	Yes	No	Free	No	Proprietary, GenMAPP

Table 1: *Overview of some popular pathway editing and analysis tools.* [a] *For pathway expansion;* [b] *Through online converter on WikiPathways.*

Another popular tools suite is Ingenuity Pathways Analysis (IPA, http://www.ingenuity.com/), an all-in-one commercial software application that enables modeling, analysis and understanding of the complex biological and chemical systems at the core of life science research. It consists of several tools enabling one to easily mine the scientific literature, build dynamic pathway models and quickly analyze high-throughput experimental data to identify key insights. IPA's dynamic pathway modeling tool applies textmining approaches for construction of novel relations between pathway entities. As a free alternative, the textmining tool Bibliosphere by the Genomatix company (http://www.genomatix.de) offers similar functionality, combining textmining results from literature with sequence data to more robustly identify relations. Bibliosphere is available from the Genomatix website. Likewise, Pathway Studio, a commercial

application suite from Ariadne Genomics, implements textmining approaches to find relationships among pathway entities. In addition, Pathway Studio can perform Gene Set Enrichment analysis on sets of genes that share a functional, biological or some other relation.

Pathway repositories

Every pathway analysis tool uses pathways that are created locally or downloaded from a central repository. To assess the quality and completeness of such repositories, we extracted and compared the pathway content from processes involved in fatty acid metabolism from several free pathway databases. We focussed on fatty acid oxidation, fatty acid synthesis and regulation of these processes. These pathways are well described in literature. Since most pathway repositories curate their pathways with literature, one would expect excellent entries. A full overview is given in **table 2**.

KEGG stores 33,679 pathways for over a hundred species generated from 269 reference pathways, about metabolism, genetic information processing, environmental response, cellular processes, human diseases and drug response. Most pathway tools make extensive use of the KEGG database. Although generally considered a robust database, some of the fatty acid metabolism pathways in KEGG have not been updated in years. The database contains the pathways for fatty acid biosynthesis, fatty acid elongation, fatty acid desaturation and fatty acid mitochondrial β-oxidation, but no entries related to the regulation of these processes.

The BioCarta pathway database offers a free collection of pathways and is hosted by a company that supplies anti-bodies for entities on several pathways. It contains key information for over 120,000 genes from multiple species through a user-friendly interface. The database contains 296 regularly updated pathways. Pathways describing mitochondrial and peroxisomal β-oxidation of fatty acids, oxidation of odd-numbered chain fatty acids, oxidation of polyunsaturated fatty acids, disease related ω-oxidation and a separate pathway of the carnitine mediated transport system are all found in BioCarta.

GenMAPP stores more than 200 contributed pathways as well as hundreds of pathways derived from the KEGG database. The contributed pathways related to human fatty acid metabolism are fatty acid synthesis, mitochondrial fatty acid β-oxidation and fatty acid degradation, but these have not been updated or checked in years.

Reactome is an online bioinformatics database of biology described in molecular terms, storing pathways as a series of separate biochemical reactions. Most pathways involved in human fatty acid metabolism are present. Some pathways, like the oxidation of unsaturated fatty acids, are not found in other databases. Several reactions, like the one converting palmitate to long-chain fatty acids, are only mentioned, whereas other databases give a complete set of reactions.

A pathway curation workflow

Biological pathway content varies greatly among the tools and databases described above in both quality and completeness, even at first glance. Additionally, pathway repositories in general apply different storage formats and conventions. Ideally, one would want to integrate all knowledge of a particular pathway from the various tools and databases available to create one 'complete' pathway. This is especially attractive when performing pathway analysis. Yet, researchers that possess the knowledge needed to create and curate such an integrated pathway are often discouraged to do so. A possible cause is the lack of a general guideline for pathway creation and curation, in addition to the false assumption that it is a time consuming process.

We ourselves apply a general biological pathway curation workflow (**figure 1**) comprised of four phases and three main data sources: online public domain databases, literature and experts. The databases used to collect content are KEGG, Reactome, GenMAPP and BioCarta, using the content from Reactome as a basis, as its content is the most complete, heavily curated and therefore of the highest quality. PubMed (http://www.pubmed.org/) is used to find relevant literature.

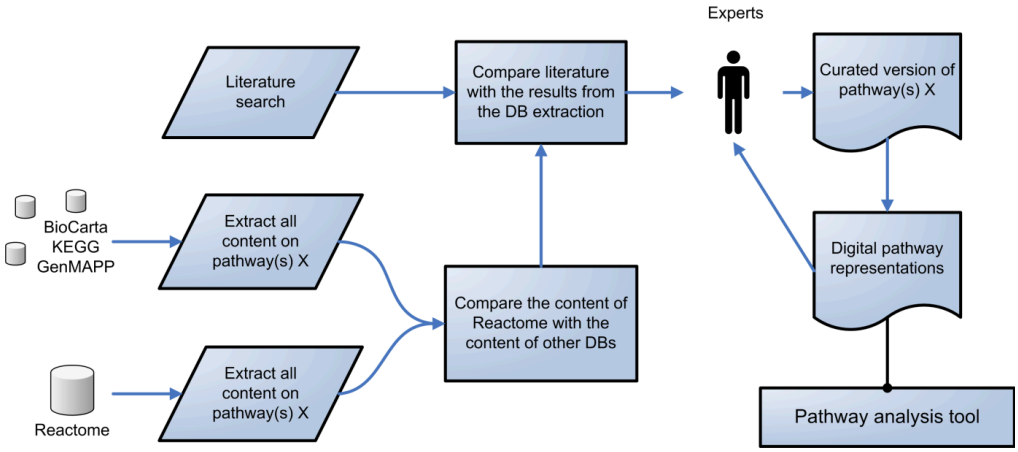


Figure 1: *Pathway curation workflow. This figure shows the outline of the biological pathway curation workflow. The workflow applies content from online databases, literature and experts. First, all biological information associated with the biological pathways in question is retrieved from Reactome and compared to content from KEGG, GenMAPP and BioCarta. Scientific literature is used to fill in the gaps and finally, the results are curated with the help of experts to remove ambiguities and inconsistencies. The curated content is used to make new digital pathway representations, suitable for analysis in a pathway analysis tool of choice. These digital representations are distributed amongst the user base of the pathway analysis tool of choice, resulting in additional curation feedback.*

The first step in this manual workflow is to collect all biological information associated with the biological pathways in question by (i) retrieving biological pathway content from several curated and highly regarded biological pathway databases, (ii) comparing and structuring content from different databases, (iii) searching for scientific literature referring to the processes described in these biological pathways and

finally (iv) curating the results with the help of experts to remove ambiguities and inconsistencies. The resulting curated content is subsequently used to create new pathways in a format suitable for a pathway analysis tool of choice, resulting in additional curation feedback.

Creating high quality curated fatty acid metabolism pathways

The manual curation process was used to create high quality curated pathways of the key processes involved in human fatty acid metabolism: fatty acid oxidation, fatty acid synthesis and regulation of these processes. Pathways involved in fatty acid metabolism are essential pathways in nutritional research. Curated up-to-date versions of these pathways are therefore in high demand.

Biological pathways in Reactome are presented as a collection of biochemical reactions and lack a visual overview displaying connections between these reactions. Hence, comparing pathway content from Reactome to the content found in other databases is a difficult task. To facilitate this process, graphical overviews were created manually by starting with the first listed reaction, connecting its product as the substrate of the second reaction and so on, creating a series of representations of all the reactions involved in human fatty acid metabolism. Once these representations were created, differences with pathway content found in the biological pathway repositories described above were annotated and simultaneously literature was searched to check everything found up to that point. Some of the pathways involved in human fatty acid metabolism were found to be incomplete or completely missing in the various biological pathway repositories. The results are best discussed on a pathway to pathway basis, highlighting the differences in completeness and accuracy between databases (summarized in **table 2**).

	Biosynthesis		Fatty acid degradation				Transport	Regulation		Fatty acid elongation
	Fatty acids	Triacyl glycerid	β ox M	β ox U	β ox P	ω ox	Carnitine shuttle	Synthesis	Degradation	Elongation
Reactome	i	c	c	c	m	m	m	i	m	m
KEGG	c	m	c	m	m	m	m	m	m	c
GenMAPP	c	m	i	i	m	m	i	m	m	m
BioCarta	m	m	c	c	i	c	c	i	m	m

Table 2: *Fatty acid metabolism pathway content of selected pathway repositories. c = complete pathway in database, m = missing in database, i = present but incomplete in database, β-ox M = mitochondrial β-oxidation, β-ox U = mitochondrial β-oxidation of unsaturated fatty acids, β-ox P = peroxisomal β-oxidation.*

Fatty acid biosynthesis has three major functions: the storage of excess energy intake, synthesis of fat from carbohydrates or proteins if the quantity of fat in the diet is low and synthesis of fat for lactation [12].

Reactome gives a good overview of the transport of citric acid from the mitochondria to the cytosol, where the citrate lygase catalyzes the production of acetyl-co-enzym A (acetyl-CoA) from citrate, to the creation of long-chain fatty acids. This pathway is followed by the triacylglycerol biosynthesis pathway.
Several essential parts of the fatty acid metabolism were found to be missing from the Reactome database. No details are given regarding the reaction that transforms palmitate into long-chain-fatty-acid and there is only a schematic overview of the reaction which transfers butyryl-acyl carrier protein (butyryl-ACP) into palmitoyl-ACP. Pathways involved in fatty acid elongation and desaturation are missing entirely.
The elongation of fatty acids proceeds through a repeated cycle of reactions. This cycle starts by the conversion of acyl-CoA to 3-ketoacyl-CoA, which is catalyzed by acetyl-CoA C-acyltransferase. The 3-ketoacyl-CoA intermediate undergoes the same three reactions that form the basis of β-oxidation, only in reverse order. Reduction of the keto-group is followed by dehydration to form a double bond. Reduction of the double bond results in an acyl-CoA that is two carbons longer than the acyl-CoA in the beginning of the cycle [13]. The desaturation takes place after elongation in the endoplasmic reticulum and is catalyzed by acyl-CoA desaturase.
KEGG describes the elongation cycles in detail and in addition gives all the enzyme codes of the different fatty acid synthases. GenMAPP has MAPPs describing fatty acid elongation and desaturation, but the reactions are not visualized in a cyclic manner.

There are several processes involved in the degradation of fatty acids, with β-oxidation being the most important [14]. B-oxidation is the process by which fatty acids are degraded in the mitochondria. The carnitine shuttle [15] is essentially the first step of mitochondrial fatty acid β-oxidation. It is involved in transport of long-chain fatty acids through both mitochondrial membranes, from the cytosol to the mitochondrial matrix where β-oxidation takes place [16].
Collecting and comparing the information on the several fatty acid degradation processes showed that Reactome does not contain any information on reactions related to the fatty acid carnitine transport system (the transport of fatty acid into mitochondria) or fatty acid cell transport system (the transport of fatty acid into the cell). The KEGG database refers to the carnitine O-palmitotransferase enzyme in the β-oxidation pathway, but fails to clarify the uptake into mitochondria. GenMAPP and BioCarta give a full overview of the carnitine shuttle, but disagree on some minor details concerning the sub-cellular location of each step in the process. These ambiguities were clarified using literature [15, 16].
Mitochondrial fatty acid β-oxidation of saturated and unsaturated fatty acids is well described in Reactome. In KEGG, β-oxidation of unsaturated fatty acids is missing. GenMAPP gives an abbreviated description and BioCarta gives a small overview. Both BioCarta and GenMAPP do not detail the seven different cycles of the β-oxidation process, falsely assuming a reiteration of one general cycle.

Peroxisomal β-oxidation [17] applies the same mechanism as mitochondrial β-oxidation, but the peroxisomal fatty acid β-oxidation system is only able to shorten fatty acids chains and cannot degrade fatty acids to completion. The shorter chains are transported as carnitine-ester from the peroxisomes to the mitochondria, where the degradation is completed. Specific information on peroxisomal β-oxidation is missing from most databases. Reactions describing the process are present, but incomplete in the Reactome database. BioCarta gives a short but otherwise complete overview.

Another form of fatty acid oxidation is ω-oxidation [17]. Ω-oxidation is a minor process that takes place in the endoplasmic reticulum, but only occurs when the β-oxidation processes are somehow impaired by disease or fasting. Information on ω-oxidation is only found in BioCarta. Ω-oxidation was not mentioned in any other database. Additions were found in literature [17].

There are five main regulatory proteins involved in the regulation of fatty acid synthesis: liver X receptor/retinoid X receptor (LXR-α/RXR-α) heterodimer, nuclear transcription factor Y (NF-Y), sterol regulatory element binding protein 1 and 2 (SREBP1, SREBP2) and carbohydrate regulatory element binding protein (ChREBP) [21-26]. The main regulator in the regulation of β-oxidation is peroxisome proliferator activated receptor alpha (PPAR-α) [18-20].
Regulation of fatty acid synthesis by ChREBP [21-26] has an entry in the Reactome database, but transcriptional activation of the synthesis via SREBP1 and LXR-α/RXR-α heterodimer and NF-Y is absent. The regulation of fatty acid degradation is not mentioned in Reactome.
BioCarta contains pathways describing regulation of fatty acid synthesis by SREBP1 and SREBP2 and the regulation of genes involved in fatty acid β-oxidation by PPAR-α. LXR is only mentioned in conjunction with FXR in a pathway on the regulation of cholesterol metabolism. Likewise, RXR degradation is mentioned, but there is no entry describing the relation to the regulation of fatty acid synthesis.
GenMAPP and KEGG do not contain any pathways describing regulation of processes involved in human fatty acid metabolism. Content from Reactome and BioCarta was expanded with the help of literature [18-26].

After all content was collected, it was passed on to experts in the field of fatty acid metabolism. New pathways describing fatty acid synthesis, triacylglyceride synthesis, fatty acid β-oxidation (saturated and unsaturated) and fatty acid ω-oxidation were created from this curated content. Pathways are available from WikiPathways (http://www.wikipathways.org/). Downloading the pathways from this location ensures that you have the most up-to-date version. WikiPathways has the option to export to formats suitable for analysis in Cytoscape, EU.gene Analyzer and GenMAPP. This resulted in constructive feedback from the large user base of these tools.

Conclusion

Biological pathway databases are far from comprehensive. We have used a rapid pathway curation workflow to collect all content of biological pathways involved in human fatty acid metabolism. At the time of writing, these improved fatty acid metabolism pathways are among the most widely used biological pathways in GenMAPP and as such have become part of the standard MAPP archive. The suggested workflow can be seen as a general guideline for anyone looking to create novel high quality curated biological pathways. The choice for a particular pathway editing tool however, is up to the user. Additionally, we use public domain databases only. Although content that comes with commercial tools is often derived from public databases and literature, those tools are in many cases well developed and have convenient user interfaces, which makes their usage beneficial for pathway developers that do have access to such tools.

A downside of the presented manual curation workflow is that it will generally not yield novel pathways when using database content as the starting point. This can be addressed by consulting experts first, which in most cases will result in a 'raw pathway'. Literature and database content can then be used to polish this raw pathway. Finding relevant literature is an arduous task, however, as querying articles for single reactions yields excessive amounts of hits. A possible solution for this issue is text mining, which can be used to generate an exhaustive list of relevant literature and networks of possible connections between queried components, for instance based on co-citation in Pubmed abstracts. These networks of inferred relations are then integrated into one or several pathways. The content should then still be curated by experts, as textmining will in general yield a large amount of false positives. In a similar fashion to manually extracted content from databases and literature, this content may then be transcribed into novel pathways. Textmining is already used in several pathway analysis tools (**see table 1**) as an alternative to pathway analysis using known curated pathways.

Using the presented workflow, it is possible to create a small collection of curated pathways within 20 man-hours, of course depending on expert availability. This is not fast enough to fill a repository with all known pathways. But this is not the purpose of the presented workflow. Anyone can make a small pathway in half an hour, but this does not imply that the information contained within the pathway is of high quality or even correct. Curation is the key. Pathway curation can normally take months, leading to out-of-date and incomplete content. The presented pathway curation workflow, although possibly less thorough, is faster and founded on the principle of public demand. Hence, it is a suitable workflow for creating key pathways. A community effort is needed to create a more robust set of pathways in a similar manner, as demonstrated by the Science Signaling and WikiPathways initiatives.

Still, to improve speed, partial automation of the curation process could be implemented, starting with content extraction from Reactome and ending at novel pathways in a suitable exchange format. There are two data flows at the moment, an overview of which can be seen in **figure 2**. Once the content is in

GPML format, the pathways are easily converted to Cytoscape, EU.gene Analyzer and GenMAPP formats through the WikiPathways interface.

Only one connection between Reactome and GPML is readily available, implementing an intermediary EBI created comma-separated text format. There is another suggested connection through the BioPAX level 2 format [27]. BioPAX Level 2 is a suitable exchange format for biological pathway content, but at the moment does not store the additional graphical layout information that is required for transparent conversion of pathway representations from pathway analysis tools such as GenMAPP and Cytoscape. Reactome offers the option to export all its pathways and reactions as BioPAX level 2 compliant files, which can then be directly imported into the java-based Cytoscape application and converted to GPML.

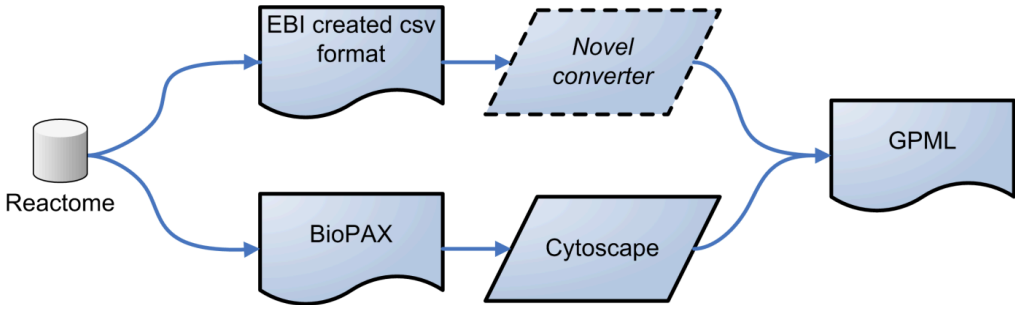


Figure 2: *Reactome to GPML pipeline. This figure shows two possible connections between Reactome and GPML. Each wave block is a data type; each parallelogram is a software application. One connection uses an EBI created comma-separated text format and a novel converter to create a GPML file, a format suitable for both WikiPathways and GenMAPP. Additionally, Reactome offers the option to export all its pathways and reactions as BioPAX level 2 compliant files, which can then be directly imported in the java-based Cytoscape application and converted to GPML.*

Pathway curation is a Sisyphean task, as new discoveries constantly lead to novel additions and adaptations to a pathway. Also, the 'beginning' and 'end' of a pathway are arbitrary definitions. This latter difficulty can be overcome by creating so-called meta-pathways, covering several separately defined pathways and their respective biochemical and biological connections, similar to the Reactome sky map. The former challenge can be met by adaptation of standard exchange formats for biological pathways, enabling automated pathway data acquisition and comparison in a network analysis tool. With such additions in place, the road to curated pathways will become less long and winding.

Acknowledgements

We would like to thank Sander Kersten and Jochum Plat, who we consulted as experts in the field of fatty acid metabolism, and Marjan van Erk, Rachel van Haaften, Peter d'Eustachio and Bernard de Bono for their valuable advice and support.

References

1. Cary, M.P. *et al*. (2005) Pathway information for systems biology. FEBS Lett, 579 (8): 1815-20.
2. Bader, G.D. *et al*. (2006) Pathguide: a pathway resource list. Nucleic Acids Res 34(Database issue): D504-6.
3. Galperin, M.Y. (2008) The Molecular Biology Database Collection: 2008 update. Nucleic Acids Res 36 (Database issue), D2.
4. Vastrik, I. *et al*. (2007) Reactome: a knowledge base of biologic pathways and processes. Genome Biol 8 (3), R39.
5. Kanehisa, M. *et al* (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res Vol. 36(Database issue), D480-D484.
6. Ekins, S. *et al*. (2007) Pathway mapping tools for analysis of high content data. Methods Mol Biol 356, 319-50.
7. Salomonis, N. *et al*. (2007) GenMAPP 2: New Features and Resources for Pathway Analysis. BMC Bioinform. 8, 217.
8. Cavalieri, D. and De Filippo, C. (2005) Bioinformatic methods for integrating whole-genome expression results into cellular networks. Drug Discov Today 10(10), 727-34.
9. Doniger, S.W. *et al*. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol, 2003. 4(1): p. R7.
10. Cline, M.S. *et al*. (2007) Integration of biological networks and gene expression data using Cytoscape. Nat Protoc 2 (10), 2366-82.
11. Pandey, R. *et al*. (2004) Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. Bioinformatics 20(13): p. 2156-8.
12. Kuhajda, F.P. (2000) Fatty-acid synthase and human cancer: new perspectives on its role in tumor biology. Nutrition 16(3), 202-208.
13. Garrett, R.H. and Grisham, C.M., eds (1999). *Biochemistry, 2nd edition*. Brooks & Cole.
14. Vockley, J. and Whiteman, A.H. (2002) Defects of mitochondrial β-oxidation: a growing group of disorders. Neuromuscular Disorders 12(3), 235-246.
15. Rubio-Gozalbo, M.E. *et al*. (2004) Carnitine-acylcarnitine translocase deficiency, clinical biochemical and genetic aspects. Molecular Aspects of Medicine 25(5-6), 521-532.
16. McClelland, G.B. (2004) Fat to the fire: the regulation of fatty acid oxidation with exercise and environmental stress. Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology 139(3), 443-460.
17. Wanders, R.J.A. (2004) Peroxisomes, fatty acid metabolism, and peroxisomal disorders. Molecular Genetics and Metabolism 83(1-2), 16-27.
18. Kota, B.P. *et al*. (2005) An overview on biological mechanisms of PPARs. Pharmacological Research 51, 85-94.

19. Mandard, S. *et al*. (2004) Peroxisome proliferator-activated receptor alpha target genes. Cell. Mol. Life Sci. 61, 393–416.
20. Rastinejad, F. (2001) Retinoid X receptor and its partners in the nuclear receptor family. Current Opinion in Structural Biology Volume 11(1), 33-38.
21. Joseph, S.B. and Tontonoz, P. (2003) LXRs: new therapeutic targets in atherosclerosis? Current Opinion in Pharmacology 3(2), 192-197.
22. Dentin, R. *et al*. (2005) Carbohydrate responsive element binding protein (ChREBP) and sterol regulatory element binding protein-1c (SREBP-1c): two key regulators of glucose metabolism and fatty acid synthesis in liver. Biochimie 87, 81–86.
23. Uyeda, K. *et al*. (2002) Carbohydrate responsive element-binding protein (ChREBP): a key regulator of glucose metabolism and fat storage. Biochemical Pharmacology 63, 2075-2080.
24. Matuoka, K. and Chen, K.Y. (2002) Transcriptional regulation of cellular ageing by the CCAAT box-binding factor CBF/NF-Y. Ageing Research Reviews 1(4), 639-651.
25. Brown, M.S. and Goldstein, J.L. (1997) The SREBP Pathway: Regulation of Cholesterol Metabolism by Proteolysis of a Membrane-Bound Transcription. Cell 89 3(2), 331-340.
26. Edwards, P.A. *et al*. (2000) Regulation of gene expression by SREBP and SCAP. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Fatty acids 1-3, 103-113.
27. Strömbäck, L. and Lambrix, P. (2005) Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. Bioinformatics 21 (24), 4401-7.

Chapter 6

Identification of novel ER-α target genes in breast cancer cells: gene- and cell-selective co-regulator recruitment at target promoters determine response to 17β-estradiol and tamoxifen

Andrea Romano[1,2,a], Michiel Adriaens[3], Sabine Kuenen[1,2], Bert Delvoux[1,2], Gerard Dunselman[1,2], Chris Evelo[3], Patrick Groothuis[1,2,4]

[1]*GROW, School for Oncology and Developmental Biology, Maastricht University, Maastricht, The Netherlands.*

[2] *Department of Obstetrics and Gynaecology, Maastricht University, Maastricht, The Netherlands.*

[3] *Department of Bioinformatics – BiGCaT, Maastricht University, Maastricht, The Netherlands.*

***Keywords**: 17β-Estradiol, estrogen receptor-α, co-activators, co-repressors, tamoxifen.*

Publication*:* Romano, A., M.E. Adriaens, S. Kuenen, B. Delvoux, G. Dunselman, C.T.A. Evelo and P. Groothuis (2010). "Identification of novel ER-alpha target genes in breast cancer cells: gene- and cell-selective co-regulator recruitment at target promoters determines the response to 17beta-estradiol and tamoxifen." Mol Cell Endocrinol 314(1): 90-100.

Abstract

Tamoxifen and 17beta-estradiol are capable of up-regulating the expression of some genes and down-regulate the expression of others simultaneously in the same cell. In addition, tamoxifen shows distinct transcriptional activities in different target tissues. To elucidate whether these events are determined by differences in the recruitment of co-regulators by activated estrogen receptor-alpha (ER-alpha) at target promoters, we applied chromatin immunoprecipitation (ChIP) with promoter microarray hybridisation in breast cancer T47D cells and identified 904 ER-alpha targets genome-wide. On a selection of newly identified targets, we show that 17beta-estradiol and tamoxifen stimulated up- or down-regulation of transcription correlates with the selective recruitment of co-activators or co-repressors, respectively. This is shown for both breast (T47D) and endometrial carcinoma cells (ECC1). Moreover, differential co-regulator recruitment also explains that tamoxifen regulates a number of genes in opposite direction in breast and endometrial cancer cells. Over-expression of co-activator SRC-1 or co-repressor SMRT is sufficient to alter the transcriptional action of tamoxifen on a number of targets. Our findings support the notion that recruitment of co-regulator at target gene promoters and their expression levels determine the effect of ER-alpha on gene expression to a large extent.

Introduction

Upon ligand activation, estrogen receptor-α (ER-α) binds to the promoters of responsive genes, interacting directly with estrogen response elements (EREs) or indirectly *via* interactions with other transcription factors (reviewed in: [1]). Numerous mechanisms participate in the fine-tuning of estrogen regulatory actions in target cells. These mechanisms allow estrogens to exert opposite transcriptional actions on different genes in the same cell type, or act as agonists in one cell type and as antagonists in another cell type. The same mechanisms may also be responsible for the unwanted side effects that have been observed during the use of estrogens and selective estrogen receptor modulators (SERMs) in medical treatments. The SERM tamoxifen, for instance, acts as an ER-α antagonist by inhibiting proliferation in breast cancer cells [2, 3], but is a partial agonist in the endometrium and increases the incidence of endometrial hyperplasia and cancer [4, 5]. The same mechanisms may also explain the resistance to tamoxifen of breast cancer patients [3, 6] and the patient-dependent therapeutic efficacy of tamoxifen to treat ovarian cancer [7].

There is increasing evidence that the gene- and cell-specific actions of estrogens depend largely on the presence of co-regulators. These proteins either bridge the ER-α / target-promoter-complex with the transcriptional machinery (co-activators such as CBP, p300, SRC family) or impair it (co-repressors; SMRT, NCoR; [1, 8]. Several recent studies have indicated that the agonistic or antagonistic action of a SERM is determined by the cellular availability of co-regulators in different cell types. For example, the agonistic action of tamoxifen in endometrial cancer cells is the consequence of high expression of the co-activator SRC-1 [9]. In breast cancer cells, down-regulation of co-repressor NCoR turns tamoxifen into an inducer of proliferation and over-expression of co-activator SRC-3 (AIB1) is predictive of resistance to

tamoxifen in breast cancer patients and is associated with malignancies in the endometrium [3, 6, 10]. However, the direct effect of co-regulators on gene transcription in distinct cell types has been demonstrated for a limited number of ER-α targets only [9, 11, 12] or by means of reporter gene assays [13, 14]. In addition, it remains difficult to understand how estrogens induce the expression of specific genes and repress the expression of others in the same cell type [15-21]; [22].

In the present study, we aimed at examining whether differential co-regulator recruitment (i) determines different transcriptional actions of one ligand on distinct target genes in the same cell type and (ii) determines the opposite transcriptional regulation of the same genes in different cell types treated with the same ligand. To this end, we applied chromatin immunoprecipitation (ChIP) together with promoter DNA array hybridisation (ChIP-chip) and identified 904 ER-α target promoters in T47D breast cancer cells. On a selection of newly identified target genes, we show that the transcriptional stimulatory or inhibitory effects of 17β-estradiol or OH-tamoxifen, the active metabolite of tamoxifen, closely correlate with the recruitment of co-activators or co-repressors, respectively. Moreover, recruitment of distinct co-regulators correlates with the opposite transcriptional responses observed in T47D and endometrial cancer cells (ECC1). To further support this notion, we show that over-expression of co-activator SRC-1 intensifies OH-tamoxifen response, while over-expression of co-repressor SMRT inhibits this response, both irrespective of cell context.

Materials and Methods

Cell lines and culture. The human breast cancer cell line T47D and human endometrial cancer cell line ECC1 were purchased from the American Type Culture Collection (ATCC; Rockville, Md. USA) and maintained as described [23]. For all experiments involving hormonal stimulation, cells were cultured for five days prior to, and during the experiment in RPMI without phenol-red (Invitrogen, Life Technologies, Inc., Carlsbad, CA) supplemented with 5% hormone-stripped serum (c.c.pro GmbH, Neustadt, Germany).

Steroid hormones. 17β-estradiol and OH-tamoxifen were purchased from Sigma-Aldrich Chemie BV (Zwijndrecht, The Netherlands). ICI-164384 was a gift from Schering-Plough (Oss, The Netherlands).

RNA extraction and cDNA synthesis. RNA was isolated using the Trizol reagent (Invitrogen, Life Technologies, Inc., Carlsbad, CA) as recommended by the manufacturer. Complementary DNA (cDNA) was synthesised using the M-MLV reverse transcriptase (Invitrogen, Life Technologies, Inc., Carlsbad, CA) as described earlier [23].

Oligonucleotides. Oligonucleotides used for linear amplification of immunoprecipitated chromatin prior to ChIP-chip and used for PCR were purchased from MWG-Biotech AG (Ebersberg, Germany) and are listed in Supplemental Table S-III.

PCR and real time PCR (RT-PCR). PCR was performed with the Taq DNA polymerase (Fermentas GMBH, St Leon-Rot, Germany) as recommended by the manufacturer. Semi-quantitative PCR was performed by stopping PCR reactions every three cycles and by evaluation of band intensity on an agarose gel. RT-PCR was performed using the Syber-green ABGene system (ABGene Limited, Epsom, United Kingdom), as recommended by the manufacturer and the BioRad MyIQ apparatus.

Chromatin immunoprecipitation (ChIP). ChIP was performed as described elsewhere [23]. Briefly, T47D or ECC1 cells were grown to 80% confluence (165 cm^2 culture flasks) treated with vehicle-only (ethanol) or with ligand for 50 minutes, fixed (1% formaldehyde, 10 minutes) and scraped in 1 ml of cold PBS supplemented with Complete™ protease inhibitor (Roche, Mannheim, Germany). After cell lysis, nuclei were pelleted, lysed and chromatin was sonicated. Chromatin-protein complexes were immunoprecipitated (IP) with protein-G/A magnetic beads (Dynal, Invitrogen Life Technologies, Inc., Carlsbad, CA) and 2 μg of specific antibodies: HC-20 against ER-α, H-224 against RNA-Pol-II, C-20, N-15 and A-22 against co-activators SRC-1, p300 and CBP, and antibodies sc-1609 and H-300 against co-repressors NCoR and SMRT (Santa Cruz Biotechnology, California, USA). After IP, bead washing and reverse crosslinking, DNA was purified using the Qiaquick reaction clean-up kit (Qiagen GmbH, Hilden, Germany). Binding of the RNA-Pol-II to the *GAPDH* promoter was used as positive control of the ChIP procedure and it was assessed using primers ChIP-positive (Supplemental Table S-III). ER-α binding to the *TFF1* promoter was used as a positive control for ChIP with ER-α antibody and it was assessed using primers in Supplemental Table S-III. ChIP PCR signals were normalised with an unspecific negative control, using primers ChIP-negative (Supplemental Table S-III) that flank cytogenetic location 12p13.3 where no transcription factors bind. All additional primers used to assess ER-α and co-regulatory protein binding are listed in Supplemental Table S-III.

ChIP-chip. ChIP in T47D cells using ER-α antibody was performed as described above. Successful ChIP was confirmed by assessing ER-α binding to the promoter of the estrogen responsive gene *TFF1*. Isolated DNA fragments were subsequently subjected to a linear-amplification as follows: a) 7.5 μl of DNA were denatured, amplified with 1.5 U of Sequenase™ T7 DNA-polymerase (Invitrogen, Life Technologies, Inc., Carlsbad, CA) using primer LA-0 (Supplemental Table S-III) in the recommended buffer (1X) for 8 minutes at 37 °C. This step was repeated once. b) 15 μl of this reaction were amplified by Taq polymerase (Fermentas GMBH, St Leon-Rot, Germany) using primer LA-1 (Supplemental Table S-III) in 0.1 mM dNTPs, 1X recommended buffer, 1.5 mM $MgCl_2$ in 100 μl final volume. Aliquots (5μl) were taken at 25, 30, 35, 40 cycles to determine the number of cycles necessary to enter the exponential phase (which was determined based on the intensity of the smeared-DNA visualised on an agarose gel). A second round of amplification using the Taq polymerase was performed. Amplified DNA was purified using the Qiaquick reaction cleanup kit (Qiagen GmbH, Hilden, Germany). Enrichment of the *TFF1* promoter was confirmed at intermediate steps of the amplification and at the end of the amplification

(**Figure 1a**). This quality control guaranteed that the amplification of signals in the ChIP-DNA did not reach saturation and therefore did not result in loss of enrichment of target promoters.
Samples were generated from three independent experiments (T1, T2 and T3). In each experiment, cells were treated with 17β-estradiol or vehicle for 50 minutes. In addition, a reference pool (P) was created by pooling equal amounts of the amplified DNA from the 17β-estradiol and vehicle-treated samples of T1, T2 and T3. The ChIP-DNA fragment was labelled with Cy-5, while the input-DNA, the DNA purified from fragmented chromatin non-subjected to IP reaction and processed through the same linear-amplification as the ChIP-DNA, was labelled with Cy-3. Labelled ChIP- and input-DNA fractions from the eight samples (four treated and four untreated) were subsequently hybridised to the Nimblegen HGS17 genome build promoter microarray containing 1500 bp of promoters from 24,134 human genes. Labelling and hybridisation were performed in-house by Nimblegen (Madison, USA). The promoter regions on the array are covered by 50- to 75-mer probes with approximately 100 bp spacing. The log-ratio of Cy-5 and Cy-3 intensities was subsequently calculated to assess enrichment of specific promoters of the ChIP-DNA compared to the input-DNA, suggesting binding of ER-α. The hybridisation efficiency of the samples from experiment T3 did not meet the quality criteria and these samples were excluded from further analysis.

Statistical analysis. Two different methods were evaluated for the identification of ER-α targets. Method (i), a within-array analysis, searches for four or more probes in each 1500 bp promoter whose signals are above a specified cut-off value. This analysis was performed using the proprietary software of Nimblegen. Method (ii) is a between-array analysis, employing positive (treated replicate samples) and negative controls (vehicle-treated samples) at probe level, which was performed in the statistical programming language R. This latter method is expected to produce a statistically more robust set of potential ER-α targets. First, the log-ratio between ChIP-DNA and input-DNA intensities is calculated separately for each array. Next, all probes are ordered according to genomic location and dichotomised using a threshold around twice the estimated standard deviation of the log-ratio. Probes with log-ratio values above this threshold are designated as positive, those below the threshold negative. Next, for each array, a sliding window of a variable number of base pairs is moved over all probes, calculating a p-value for each window with a Yates corrected chi-square test. To determine whether a promoter shows true significant enrichment, the promoter has to contain at least one window that shows significant enrichment in at least two treated samples (positive controls) and the same window or windows should not show significant enrichment in more than one untreated sample (negative controls). To minimise false positives, an adaptation of the Benjamini and Hochberg method [24] is applied to calculate false discovery rates (FDR). Both methods showed over fifty percent consistency when a FDR threshold of 20 % was applied. We compared the list of target genes obtained with the two methods with a list of already known targets [25]. Given that at the same FDR, method (ii) retrieved a larger number of known target promoters when compared to method (i) and considering the greater robustness of a between-array approach, method (ii) was used to generate the list of targets used for further analysis.

To identify our 904 promoters, we combined results using two FDR cut-off points. We first identified a suitable cut-off point able to retrieve as many previously found targets [25] as possible. Using a FDR cut-off of 20 % we identified most known targets (i.e. *CTSD*, *BRCA*, *c-Myc*, *ADORA1, AGT, HSPB1*, *LCN2*) and only few more (*TGFA*, *TERT*) were retrieved when cut-off points with lower stringency (FDR cut-offs higher than 20 %) were used. Therefore, 20 % FDR was fixed as the upper limit for the stringency of our statistics. Subsequently, a low stringency (FDR 20 %) was used to identify ER-α targets common in the arrays of the independent experiments (T1 and T2 or T1, T2 and P). A high stringency (FDR 5 %) was used for targets that were common in one of the T arrays and the P array, as those areessentially technical replicates.

The promoter regions were scanned for occurrence of EREs using the Genomatix MatInspector software [26] and the Genomatix transcription factor motif database (www.genomatix.de). We also scanned promoter sequences of a validated sub selection of ER-α targets for the presence of potential tethering domains for EREs (AP1, NFκB and SP1 binding sites), using the same approach.

Cell transfection, luciferase assay and immunocytochemistry. Plasmids used for transfection were previously described: *ERE-TK-luciferase* (2X ERE-TK-LUC) containing the estrogen responsive promoter-luciferase reporter [27], was gifted by Prof Scheule. The expression vector for co-activator SRC-1 [14] and the co-repressor SMRT [28] were gifts from Prof O'Malley and Prof Evans, respectively. The SMRT expression plasmid used in these experiments encodes for a truncated form of the human co-repressor SMRT (amino-acids 1032-2517) with a dominant co-repressing action [13]. Plasmid pCNDA3.1 (Invitrogen, Life Technologies, Inc., Carlsbad, CA) was used as empty vector (when indicated). All techniques were previously described [23]. In short, transfection was performed using the jetPEI^{TM} reagent (Q-Biogene, Heidelberg, Germany) as recommended by the manufacturer. Prior to luciferase assays, cells were cultured in two wells of a 12-well plate and were transfected (2 μg DNA plus 3 μl jetPEI^{TM} per well). Sixteen hours after transfection, cells from the two wells were trypsinised, pooled and seeded into 12 wells of a 96 well-plate. Eight hours after plating, treatments were applied. Each treatment was performed in triplicate (the number of initially transfected wells was scaled up according to the number of stimulations needed). In case of RNA isolation, cells were transfected in two 25 cm^2 flasks (10 μg DNA plus 15 μl jetPEI^{TM} per flask) and subsequently cells were pooled and plated in 9 wells of a 12-well plate. For immunocytofluorescence, cells were cultured on glass cover slips fixed in buffered formaldehyde (4% paraformaldehyde in PBS), permeabilised with 0.1% Triton-X-100 in PBS and stained with the following antibodies (as indicated in the figures): goat polyclonal C-20 against co-activator SRC-1 and sc-1609 against co-repressor NCoR (Santa Cruz Biotechnology, California, USA), followed by anti-goat FITC secondary antibody 705-095-147 (Jackson Immunoresearch/Brunschwig chemie B.V., Amsterdam, The Netherlands); rabbit polyclonal H-300 against co-repressor SMRT (Santa Cruz Biotechnology, California, USA), followed by anti-rabbit FITC F005401 (DAKO, Glostrup, Denmark). For western blot (**Supplemental Figure S-1**) ER-α was detected with monoclonal antibody F10 (Santa Cruz

Biotechnology, California, USA), whereas p300 and CBP with rabbit A-22 and N-15 antibodies, respectively (Santa Cruz Biotechnology, California, USA). Mouse antibody AC-15 (Sigma-Aldrich Chemie BV, Zwijndrecht, The Netherlands) was used to detect β-actin. HRP-conjugated rabbit anti-mouse-antibodies (DAKO, Glostrup, Denmark) and goat-anti-rabbit-antibodies (Pierce, Aalst, Belgium) and the super signal-R West-Femto kit (Pierce, Aalst, Belgium) were used for primary antibody visualisation.

Results

Identification of genomic binding sites for ER-α

ER-α binding sites in gene promoters were searched genome-wide using the estrogen-responsive T47D breast cancer cells. Estrogen-responsiveness was shown by the expression of ER-α, the induction of various known estrogen responsive genes (*TFF1*, *c-Myc*, *CCND1*) and by the induction of cell proliferation by 17β-estradiol (**Supplemental Figures S-1 and S2**). T47D cells were incubated with 1 nM 17β-estradiol for 50 minutes, which was shown to result in maximal ER-α binding to the *TFF1* promoter [16, 29]; this study, results not shown). After chromatin immunoprecipitation (ChIP) using an ER-α antibody, two rounds of nucleic acid amplification were performed to yield sufficient DNA for array hybridisation to the Nimblegen promoter arrays. In order to assure adequate quality of the amplified DNA fragments, enrichment of the *TFF1* promoter was confirmed after each amplification round (**Figure 1a**). Three independent experiments, each consisting of a 17β-estradiol and a vehicle treated sample, were performed (T1, T2 and T3). Given that the hybridisation performance of the T3 samples was poor, data from experiment T3 were not used for subsequent analyses. An additional sample was included (referred to as the pool, P) created by combining equal amounts of amplified DNA material from T1, T2 and T3.

We applied robust statistical procedures (see 'Materials and Methods'), which allowed us to retrieve several previously known ER-α target promoters (i.e. *CTSD*, *BRCA*, *c-Myc*, *ADORA1, AGT, HSPB1*, *LCN2*; [25]. With this method, 904 potential ER-α binding sites were identified in total (Supplemental Table S-I), some of which are common to recent genome-wide screenings for ER-α targets (Supplemental Table S-II). The 904 binding sites are equally distributed over all chromosomes (Supplemental Table I), excluding the Y chromosome, as the T47D line is derived from a woman. Only one site was found on chromosome Y and is not included in the list of 904 targets.

ChIP-chip validation and target promoter features

To validate the findings of the ChIP-chip, standard ChIP assays were performed on additional independent experiments (two or more) and ER-α binding was confirmed for a selection of 12 promoter regions (**Figure 1b**). Enrichments were not seen for three non-target locations (*PGR* gene exons 4 and 6 and chromosome region 12p13.3).

To demonstrate that ER-α binding to the promoter regions is functional, the effect on mRNA expression was studied with RT-PCR (**Figure 1b**). The expression of most genes is induced by 17β-estradiol, with the

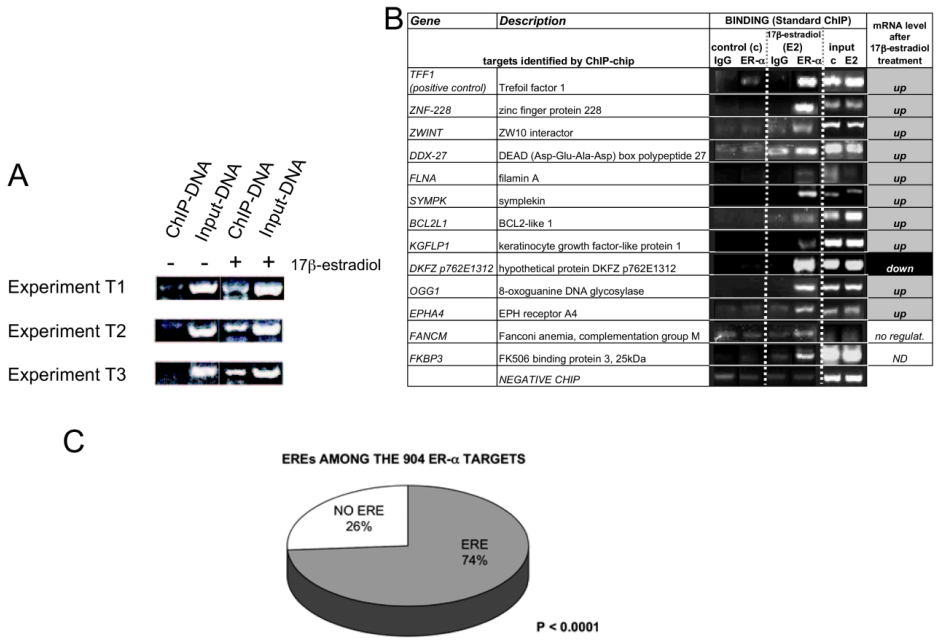


Figure 1. ChIP-chip: quality control, validation and prevalence of EREs

A. Prior to ChIP-chip hybridisation, immunoprecipitated (IP) DNA fragments were amplified (linear amplification). As a quality check, binding of ER-α to the TFF1 promoter was confirmed after each amplification round (shown for each experiment at the end of the amplification, just prior to labelling and hybridisation). ChIP-DNA = IP DNA. Input-DNA = non-IP- chromatin amplified similarly to the ChIP-DNA.

B. ER-α targets identified by ChIP-chip and validated by standard ChIP. For all ChIP experiments, cells were treated for 50 minutes; control = vehicle treated cells; E2 or 17β-estradiol: 1 nM. IgG = ChIP with non-specific antibodies; ER-α = ChIP with an ER-α antibody. Column on the right: mRNA level of the corresponding gene after 17β-estradiol (1 nM) induction. mRNA was assessed (RT-PCR or semi-quantitative PCR – semiQ-PCR) after different periods of hormone stimulation (up to 24 hours) in triplicate. Results in column signify that the considered mRNA is significantly regulated in the indicated direction ($p < 0.05$ compared to time point zero) at one time point at least (results not shown). ND: not determined.

C. Prevalence of ERE in the promoters of the entire group (n = 904) of ER-α target genes as determined by Genomatix MatInspector (http://www.genomatix.de). Promoters were scanned using a family of ERE consensus matrices [26]. A specification of EREs present in the group of 904 targets is given in the Supplementary Table S-IV.

exception of *DKFZ p762E1312*, which is down-regulated, and *FANCM*, which does not respond despite ER-α binding to its promoter (**Figure 1b**). In addition, we evaluated the transcriptional response of six target genes for which ChIP reactions were not set-up, *CCNE2*, *IGF1-R*, *FBP-1*, *BCL2*, *MALL* and *CA2* (**Supplemental Figure S-3**). All genes, except *CA2,* are induced by 17β-estradiol. *MALL* and *CA2* are induced by OH-tamoxifen, whereas *BCL2* and *CCNE2* expression is reduced by OH-tamoxifen.

Binding sites for ER-α are present both upstream and downstream of the transcription start site (TSS) and are evenly distributed along the promoter regions with respect to the distance from the TSS (results

not shown). Seventy four percent of the 904 target promoters contain an estrogen-response element (ERE; **Figure 1c**) in silico, determined with the Genomatix MatInspector software.

Selective recruitment of co-regulators determines the ER-α mediated transcription

Both 17β-estradiol and OH-tamoxifen can simultaneously up- and down-regulate the transcription of different genes in the same cell. To verify whether differential co-regulator recruitment (i.e. co-activators *versus* co-repressors) accounts for these opposite transcriptional responses in the same cells, we performed ChIP with antibodies directed against ER-α, co-activators p300, CBP and SRC-1 or co-repressors SMRT and NCoR after exposing T47D cells for 50 minutes to 1 nM 17β-estradiol or to 1 μM OH-tamoxifen. These co-regulators were selected because they are expressed in T47D cells (**Supplemental Figure S-1**) and all three co-activators are efficiently recruited at the promoter of *TFF1* after 17β-estradiol induction (results not shown). It should be noted that we did not aim at identifying which specific co-regulator binds to one region, but rather whether co-activators or co-repressors are recruited. CBP / p300 are general mediators, which bridge the basal transcriptional machinery to the ER-α complex with additional co-activators, irrespective to which specific protein is present (SRC1, SRC2 or SRC3; [30, 31]. Therefore, in order to immunoprecipitate all DNA sequences interacting with co-activators simultaneously, we pooled the antibodies against p300, CBP and SRC-1. For the same reasons, we pooled co-repressors NCoR and SMRT antibodies.

The expression of *TFF1*, *DDX-27*, *ZNF-228* and *ZWINT* is up-regulated by 17β-estradiol and down-regulated by OH-tamoxifen (**Figure 2**), which correlates well with the recruitment of co-activators and co-repressors, respectively. In contrast, the expression of *FLNA*, *SYMPK*, *KGFLP1* and *BCL2L1* is induced by both 17β-estradiol and OH-tamoxifen (**Figure 3**). In these cases, recruitment of predominantly co-activators is observed, although for some gene-promoters a non-significant recruitment of co-repressors can be seen as well (*BCL2L1* after 17β-estradiol treatment and *FLNA*, *SYMPK*, *KGFLP1* after OH-tamoxifen treatment). Expression of *DKFZ p762E1312* is suppressed by both 17β-estradiol and OH-tamoxifen (**Figure 4a**). In the presence of 17β-estradiol, ER-α recruits co-repressors only; however, in the presence of OH-tamoxifen, co-activators are recruited as well (**Figure 4a**). This could be explained by the fact that OH-tamoxifen induces the transcription of *DKFZ p762E1312* at later time points (**Supplemental Figure S-3**). Also in case of the transcription up-regulation by 17β-estradiol of *EPHA4* (**Figure 4b**), ER-α recruits co-activators at the *EPHA4* promoter. No recruitment of co-regulators is observed for this gene in response to OH-tamoxifen (**Figure 4b**) and its transcription is not altered, even though ER-α binds to the promoter.

Differential recruitment of co-regulators determines cell-specific transcriptional activities of ER-α

We examined whether co-activators and co-repressors are recruited to selected ER-α target genes in accordance with their opposite transcriptional responses to OH-tamoxifen in T47D breast cells *versus* ECC1 endometrial cancer cells (ECC1 cells are ER-α / co-regulator positive - **Supplemental Figure S-1** -

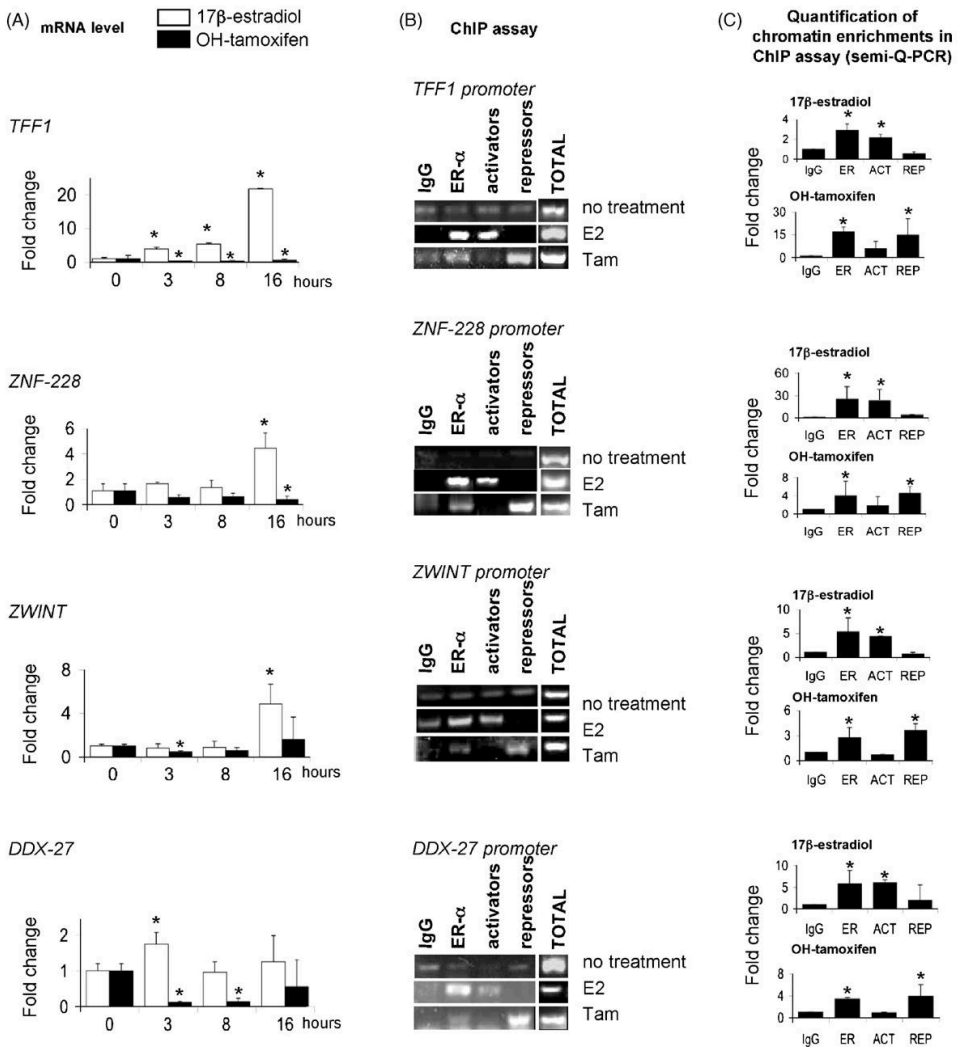


Figure 2. Co-regulator recruitment at targets induced by 17β-estradiol and repressed by OH-tamoxifen in T47D cells

A. Transcriptional responses of the indicated target genes (RT-PCR) after treatment with 17β-estradiol, OH-tamoxifen (1 nM and 1 μM, respectively) in T47D. Mean ± standard deviation (SD), n = 3. Asterisks: p < 0.05 compared to time point zero. Expression data were reconfirmed in at least one extra independent experiment.

B. ChIP assessing binding of ER-α, co-activators (p300, CBP and SRC-1) or co-repressors (SMRT and NCoR) to the corresponding promoter (E2 = 17β-estradiol. Tam = OH-tamoxifen. No treatment: treatment with vehicle only (ethanol)). Cells were treated for 50 minutes before ChIP.

C. Quantitative evaluation (estimated by agarose-gel band intensities) of chromatin enrichments after ChIP with ER-α, co-activator (ACT) or co-repressor (REP) antibodies. Mean ± SD; n = 2 or 3. Asterisks: p < 0.05 compared to the IgG control. ChIP experiments were reconfirmed in at least one additional independent experiment. The ChIP negative control for these assays is shown in Figure 4c.

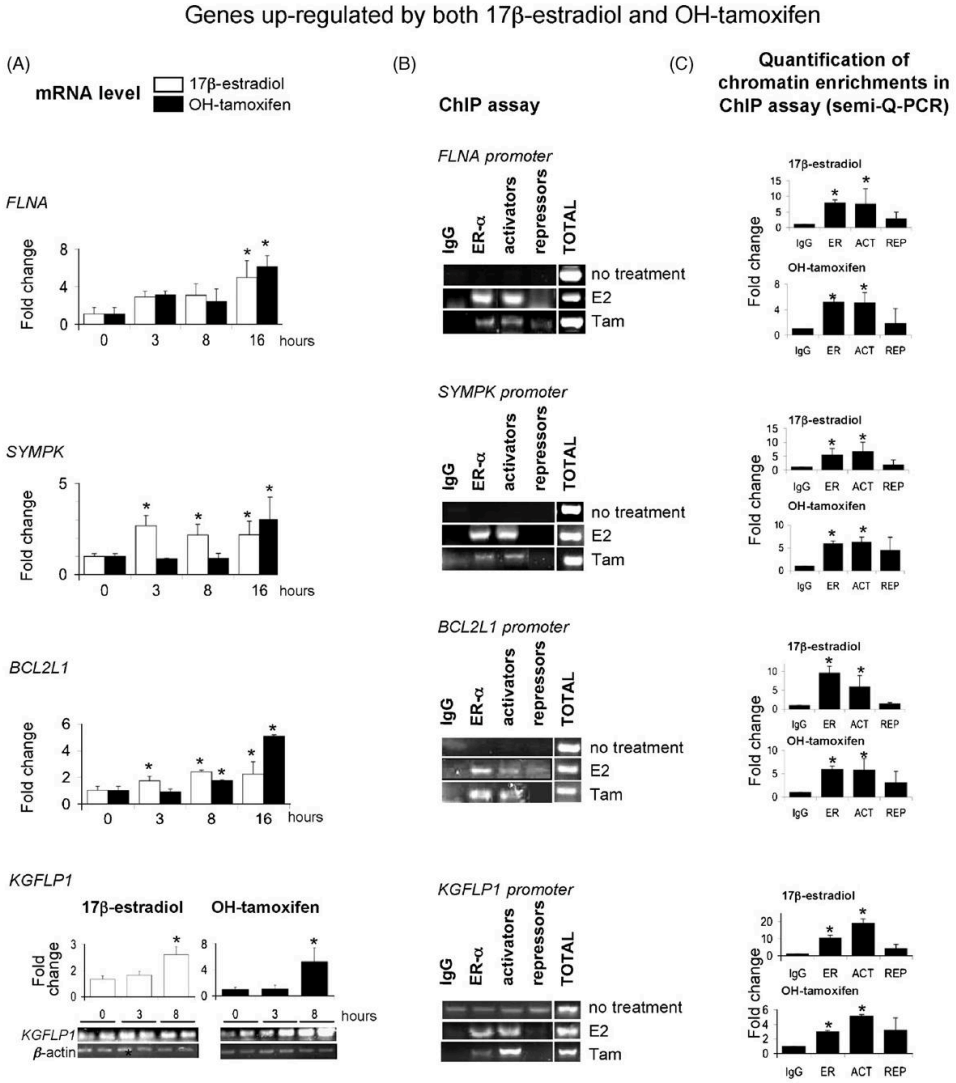


Figure 3. Co-activators are recruited at genes induced by both 17β-estradiol and OH-tamoxifen in T47D cells

A. Transcriptional responses in T47D to 1 nM 17β-estradiol or 1 μM OH-tamoxifen (RT-PCR and semiQ-PCR for KGFLP1). Mean ± SD, n = 3. Asterisks: $p < 0.05$ compared to time point zero. Expression data were reconfirmed in at least one independent experiment.

B. ChIP assessing binding to the corresponding promoter of ER-α, co-activators or co-repressors (50 minutes after induction start: E2 = 17β-estradiol. Tam = OH-tamoxifen. No treatment: induction with vehicle only).

C. Quantitative evaluation of chromatin enrichments after ChIP with ER-α, co-activator (ACT) or co-repressor (REP) antibodies. Mean ± SD; n = 2 or 3. Asterisks: $p < 0.05$ compared to IgG control. ChIP experiments were reconfirmed in at least one additional independent experiment (ChIP negative in Figure 4c).

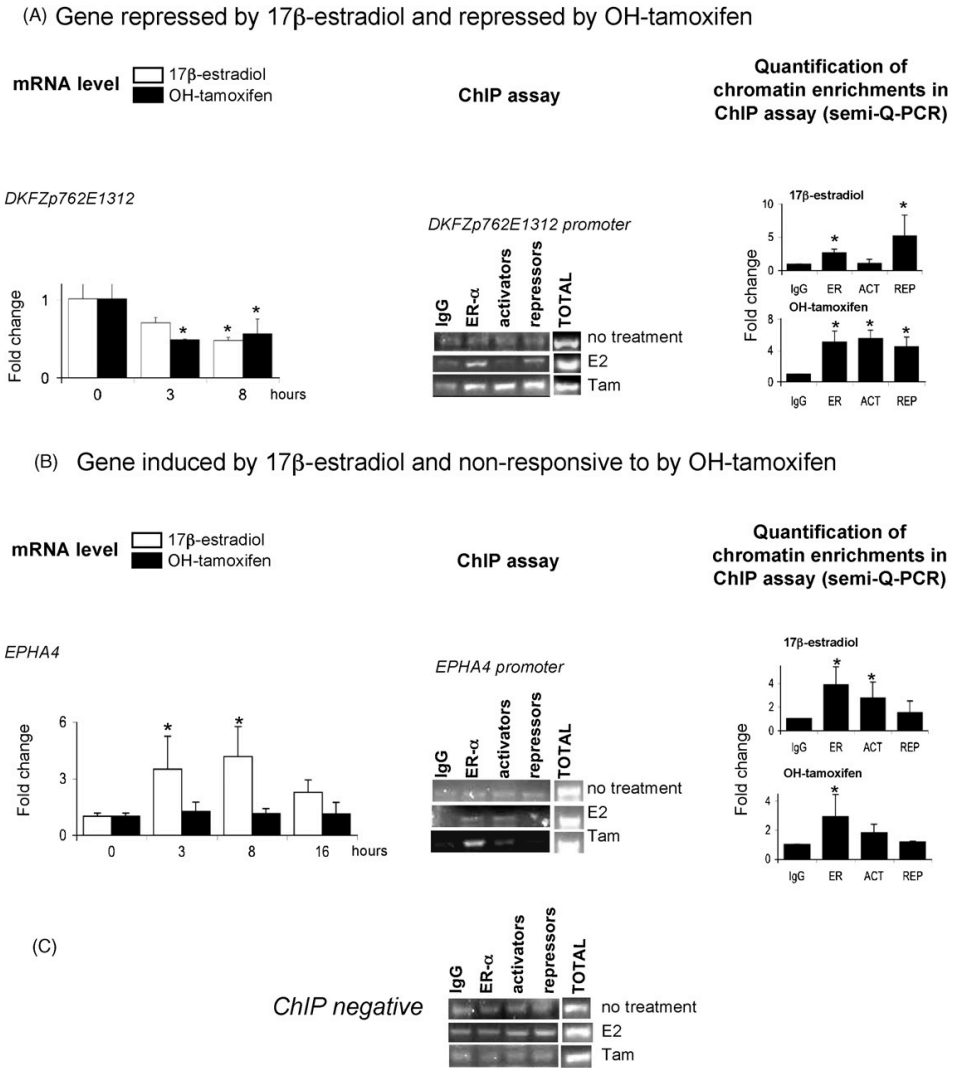


Figure 4. Co-regulator recruitment by activated ER-α at DKFZ p762E1312 and EPHA4 in T47D cells
A. On the left: transcriptional responses of the DKFZ p762E1312 gene (repressed by both 17β-estradiol - 1 nM - and OH-tamoxifen - 1 μM) in T47D (RT-PCR). Mean ± SD, n = 3. Asterisks: p < 0.05 versus time point zero. RNA data were reconfirmed in at least one extra independent experiment. Middle: ChIP assessing binding to the DKFZ p762E1312 promoter of ER-α, co-activators or co-repressors. ChIP was performed 50 minutes after induction start: E2 = 17β-estradiol. Tam = OH-tamoxifen. No treatment: vehicle only. Right: quantitative evaluation of chromatin enrichments after ChIP with ER-α, co-activator (ACT) or co-repressor (REP) antibodies. Mean ± SD. Asterisks: p < 0.05 versus IgG control, n = 2 or 3.
B. EPHA4 gene is induced by 17β-estradiol (1 nM) but is not influenced by 1 μM OH-tamoxifen (on the left; mean ± SD, n = 3. Asterisks: p < 0.05 versus time point zero). RNA data were reconfirmed in at least one extra independent experiment. Middle and right: ChIP assay and quantitative evaluation of the ChIP experiments (mean ± SD based on at least two independent experiments. Asterisks: p < 0.05 versus IgG control). C. ChIP negative control (cytogenetic location 12p13.3).

and estrogen-responsive - **Supplemental Figure S-2**). In ECC1 cells, *KGFLP1, DDX-27* and *FLNA* are induced by 17β-estradiol and OH-tamoxifen, whereas *TFF1* is induced by 17β-estradiol only. ER-α preferentially recruits co-activators to up-regulate these genes (**Figure 5**). In contrast, the transcriptional inhibitory effects of 17β-estradiol (for *BCL2L1*) or OH-tamoxifen (for *TFF1*, *BCL2L1* and *EPHA4*) are associated with the recruitment of co-repressors after ER-α binding (**Figure 5**).

Interestingly, OH-tamoxifen and 17β-estradiol reduce the expression of *BCL2L1* in ECC1, but induce it in T47D cells (**Figures 5b and 3**, respectively). In contrast, the expression of *DDX-27* is induced in ECC1 and reduced in T47D cells by OH-tamoxifen (**Figures 5b and 2**, respectively). These opposite transcriptional effects are clearly related to the recruitment of different co-regulatory proteins in the two cell contexts: co-activators in case of induction, and co-repressors in the case of inhibition of transcription. The same is evident for *EPHA4*. This gene is induced by 17β-estradiol in T47D cells, under which condition ER-α recruits co-activators (**Figure 4b**). However, *EPHA4* is not responsive to 17β-estradiol in ECC1 cells, and in this cell context, binding of ER-α to the corresponding promoter is not accompanied by further co-regulator recruitment (**Figure 5b**). The opposite is observed with OH-tamoxifen, which inhibits *EPHA4* expression in ECC1 cells but has no effect T47D cells. In T47D cells, no co-regulators are recruited by ER-α (**Figure 4b**), whereas in ECC-1 cells, binding of ER-α is followed by recruitment of co-repressors (**Figure 5b**). The recruitment of distinct co-regulators at the promoters of *DDX-27* and *BCL2L1* in T47D and ECC1 after induction with OH-tamoxifen was confirmed by real-time PCR (**Figure 5d**).

Over-expression of SRC-1 and SMRT alters the response of target genes to OH-tamoxifen

If the regulation of the aforementioned genes is truly dependent on co-regulators, it should be expected that, as previously shown [13, 14], modification in the level of some of these proteins modifies the response of the target genes. Therefore, to confirm the association between up- or down-regulation and recruitment of co-activators or repressors, we over-expressed co-activator SRC-1 or co-repressor SMRT by transient transfections in T47D and ECC1 cells (**Figure 6a**). To proof that these transfections had significant and measurable effects, we assessed the activity of the estrogen-responsive construct *ERE-TK-luciferase* after co-transfection with SRC-1 or with SMRT. As expected, SRC-1 over-expression enhances the 17β-estradiol-induced luciferase activity, whereas SMRT reduces it (**Figure 6b**). To confirm the transfectability of T47D and ECC1 cells we also measured GFP expression after transient transfection with a GFP expression plasmid (**Supplemental Figure S-4**).

Figure 6c shows the effect of SRC-1 or SMRT over-expression on a number of identified target genes. In T47D cells, *BCL2L1* transcription is normally up-regulated by OH-tamoxifen. Over-expression of the co-activator SRC-1 enhances this effect, whereas over-expression of the co-repressor SMRT changes OH-tamoxifen into an inhibitor of transcription (**Figure 6c**). In ECC1, *BCL2L1* is normally repressed by OH-tamoxifen, but over-expression of SRC-1 changes OH-tamoxifen into an inducer of transcription.

With regard to the expression of *EPHA4*, over-expression of SRC-1 in T47D cells turns OH-tamoxifen into an inducer of transcription, whereas this gene is unresponsive under normal conditions. In ECC1 cells,

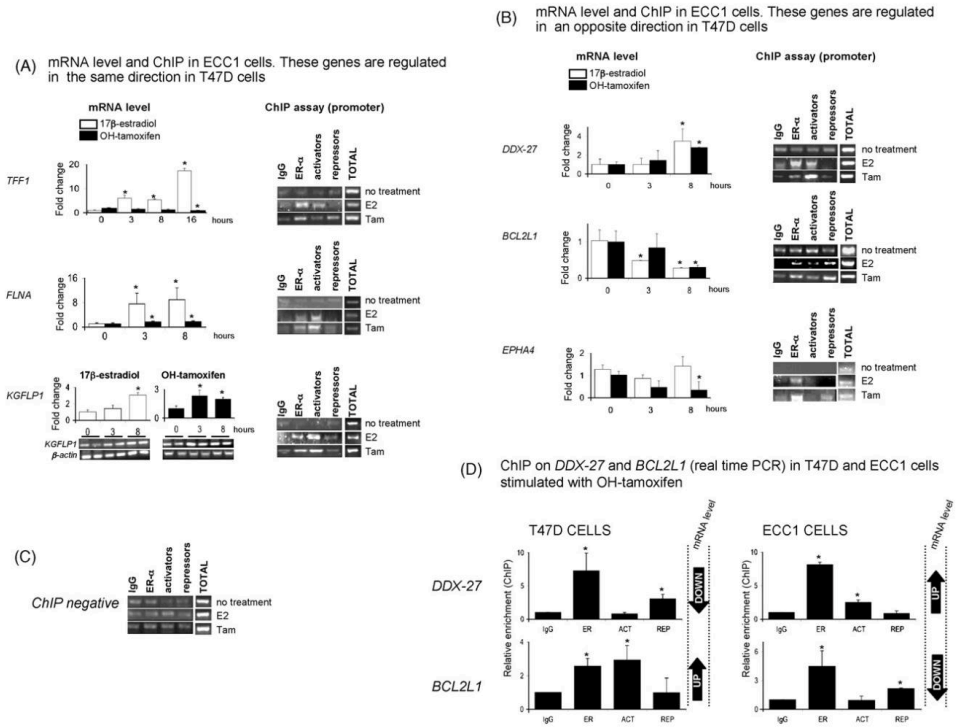


Figure 5. mRNA level and co-regulator recruitment in ECC1 cells

A and B. Transcriptional responses (RT-PCR and semiQ-PCR for KGFLP1) after 17β-estradiol or OH-tamoxifen stimulation (1 nM and 1 μM, respectively) in ECC1 (left side of panels A and B). Mean ± SD, n = 3. Asterisks: $p < 0.05$ versus time point zero. RNA expression data were reconfirmed in at least one additional independent experiment. ChIP assays (50 minutes of induction) showing binding of ER-α, co-activators (SRC-1, CBP and p300) and co-repressors (NCoR and SMRT) to the corresponding promoter are shown on the right of each A and B panels (E2 = 17β-estradiol. Tam = OH-tamoxifen. No treatment: vehicle only). ChIP experiments were reconfirmed in at least one additional independent experiment.

A. The transcriptional response of these genes in ECC1 (shown in panel) is similar to the response observed in T47D cells (shown in Figures 2a and 3a) and ChIP indicates that the same kind of co-regulators are recruited at gene promoters in the two cell lines (ECC1, shown in this figure, and T47D cells, Figures 2 and 3).

B. The transcriptional response of these genes in ECC1 (shown in panel) is opposite compared to the response observed in T47D cells (shown in Figures 2, 3 and 4) and ChIP indicates that the distinct co-regulators are recruited at gene promoters in ECC1 (shown in panel) and T47D cells (Figures 2, 3 and 4).

C. ChIP negative control (cytogenetic location 12p13.3).

D. Relative enrichments of ER-α, co-activators (ACT) or co-repressors (REP) at the promoters of DDX-27 and BCL2L1 in T47D and ECC1 cells after OH-tamoxifen induction (50 minutes). OH-tamoxifen induces DDX-27 and BCL2L1 in opposite directions in T47D and ECC1 cells. The direction of the mRNA regulation is indicated by the arrows. ChIP reactions were measured by real-time PCR (mean ± SD based on two replicates. Asterisks: $p < 0.05$ versus IgG control).

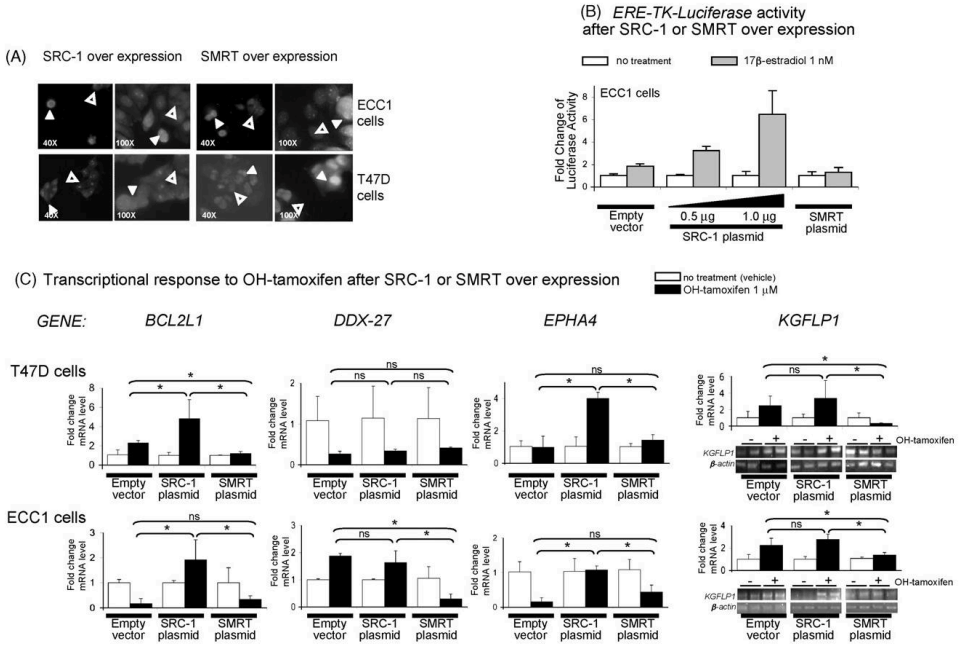


Figure 6. Over-expression of SRC-1 and SMRT modifies OH-tamoxifen responses

A. Over-expression of co-activator SRC-1 and co-repressor SMRT in T47D and ECC1 cells after transient transfection (immunocytofluorescence). Empty arrow-heads: endogenous expression level. Solid arrow-heads: over-expressing cells.

B. Induction of the ERE-TK promoter after co-transfection of ECC1 cells with the 2X ERE-TK-LUC construct (containing the luciferase reporter) along with either the expression plasmid for co-activator SRC-1 (increasing amounts of plasmids used for transfection) or the plasmid expressing co-repressor SMRT. Cells were transfected as described in material and methods in 12-well plates using 2 μg of total plasmid DNA: 1 μg of 2X ERE-TK-LUC combined with variable amounts (0 – 1 μg of SRC-1). Total amount of transfected DNA was kept constant using the empty vector. For induction (n = 3 per treatment ± SD) and luciferase assay, transfected cells were re-plated on a 96 well-plate. Similar results are obtained in T47D cells (not shown).

C. Transcriptional responses of BCL2L1, DDX-27, EPHA4 (RT-PCR) and KGFLP1 (semiQ-PCR) after stimulation with 1 μM OH-tamoxifen or with vehicle only (no treatment) for 5 hours in T47D and ECC1 cells transiently transfected with the empty vector, SRC-1 expression plasmid or SMRT expression plasmid. Cells were transfected as described in material and methods in 25 cm^2 flask (10 μg DNA) and re-plated for induction and RNA isolation in 12-well plates. Bars indicate mean ± SD, n = 3.

EPHA4 transcription is inhibited by OH-tamoxifen and SRC-1 over-expression impairs this repressive activity. Also in case of the transcriptional activation of *KGFLP1* in both T47D and ECC1 cells, SMRT over-expression is sufficient to reverse (in T47D cells) or impair (in ECC1 cells) this response (**Figure 6c**). Transcription of *DDX-27* is suppressed by OH-tamoxifen in T47D and induced in ECC1 cells. Over-expression of SRC-1 does not affect the inhibitory action of OH-tamoxifen in T47D, but over-expression of SMRT in ECC1 cells turns OH-tamoxifen into a repressor of transcription (**Figure 6c**).

The response to OH-tamoxifen of other validated genes (*TFF1*, *FLNA*, *SYMPK*, *DFFZ p762E1312*, *ZWINT* and *ZNF-228*) and the responses to 17β-estradiol in general, were not significantly influenced by modifications of the level of SRC-1 and SMRT (data not shown), suggesting either a promoter-specificity or a cell-specific modification of co-regulators as a mechanism behind the distinctive interaction with the gene promoters of these genes.

Discussion

The aim of the present study was to elucidate the role of co-regulators in (i) the opposite transcriptional actions mediated by ER-α on different target genes and (ii) the tissue specific actions of OH-tamoxifen (and 17β-estradiol) in breast and endometrial cancer cells. To this end, we first identified ER-α target promoters genome-wide by ChIP-chip and subsequently we examined whether co-activators or co-repressors are recruited by activated ER-α at the promoters of a number of newly-identified targets.

Though some past studies have focussed on the genome-wide identification of ER-α binding sites in breast cancer cell lines [15, 16, 18-21, 29, 32-34] none have further considered the role of co-regulators on the transcriptional regulation of these ER-α targets. Up to now, this knowledge has been generated by means of reporter gene assays [13, 14] or by studying a low number of estrogen responsive genes only [9, 11].

In the present study, we identified 904 promoters targeted by ER-α using ChIP-chip. These results were validated by standard ChIP, by the estrogen responsiveness of the corresponding genes at the mRNA level, and by the high prevalence of EREs among target promoters (**Figure 1**).

Co-regulator recruitment at target promoters determines gene- and cell-specific responses to ER-α ligands

In line with previous studies [9, 11, 12], activated ER-α binds to gene promoters, recruits co-activators or co-repressors, which determine the subsequent transcriptional up- or down-regulation, respectively (**Figures 2, 3 and 4**). In one cell type, all determinants of the ER-α action (like ligand concentration, level and activation of ER-α and co-regulators) are identical, except for the promoter, which therefore must be responsible for the recruitment of different co-regulators. A number of studies have already shed light on the roles of ERE-motifs and additional cis-regulatory elements (AP1, Sp1, NFκB binding sites) in the cell- and ligand-specific regulation of ER-α and ER-β [35-37]. The main features of the genes analysed in the present study (EREs and binding sites for additional transcription factors) are given in Supplemental Table S-IV. Alternatively, it is possible that co-regulators are modified post-translationally in a cell-specific manner, resulting in altered interactions at gene promoters in the distinct cell contexts.

In one case only (*DKFZ p762E1312*), transcription repression by OH-tamoxifen was associated with recruitment of both co-repressors and co-activators. We explained this effect with the ability of OH-tamoxifen to induce *DKFZ p762E131* transcription at later time points. However, it should also be noted that the dynamics, the sequential and combinatorial assembly of co-activators and co-repressors at target

promoters have not been addressed in the present investigation. Nevertheless, these events are important for the action of nuclear receptors [38, 39].

Differential co-regulator recruitment also explains the opposite transcriptional response observed at a number of target genes in response to OH-tamoxifen (*DDX-27*, *BCL2L1* and *EPHA4*) or 17β-estradiol (*BCL2L1* and *EPHA4*) in breast cancer (T47D; **Figures 2, 3 and 4**) and endometrial cancer cells (ECC1; **Figure 5**). These results confirm a previous finding for a number of known estrogen responsive genes (*c-Myc*, *IGF-I*, *EBAG9* and *CTSD*; [9]. The present study extends this mechanism of action to potentially all ER-α target genes.

To further substantiate the association between transcriptional regulation and co-regulator recruitment, we over-expressed either co-activator SRC-1 or co-repressor SMRT. In a number of cases, the transcriptional response to OH-tamoxifen in T47D or ECC1 cells could be modified or inverted by over-expression of these co-regulators (*BCL2L1*, *KGFLP1*, *EPHA4*; **Figure 6**).
The transcription of other genes in response to OH-tamoxifen was not influenced by SRC-1 or SMRT over-expression (*TFF1*, *FLNA*, *SYMPK*, *DFFZ p762E1312*, *ZWINT* and *ZNF-228*). In some cases, as observed for *DDX-27*, the inducing action of OH-tamoxifen could be impaired in ECC1 after over-expression of SMRT, but the opposite inhibitory action of OH-tamoxifen observed in T47D cells could not be changed by SRC-1 over-expression. As shown by others [13, 40], each promoter interacts with a limited number of co-regulators only and therefore each co-regulator modulates the expression of a limited number of genes. These events explain why co-regulators have distinct physiological functions [41-44]. In our case, it entirely possible that SRC-1 cannot be efficiently recruited at the *DDX-27* promoter, whereas neither SRC-1 nor SMRT can be efficiently recruited at the promoter of other target genes, whose transcription was not influenced by these two co-regulators.

Conclusions

Complex events determine the action of ER-α, including histone modifications [45], distal and proximal cis-regulatory elements [16], ligand independent signalling and indirect DNA binding mediated by additional transcription factors. Our results suggest that at least for direct ER-α targets distinct co-regulator recruitment is one of the key modulators of hormonal response.
In case of important drugs like tamoxifen, ER-α is necessary but not sufficient to mediate its actions. The direction of the hormonal response is for a large part dependent on co-regulators. Aberrations in the functions mediated by these proteins may lead to endocrine related cancers, to innate and developed drug-resistance in breast tumours [3, 6, 10] or to poor therapeutic response of ovarian tumours [7]. Unravelling the expression and activation patterns of co-regulators in estrogen-dependent tumours may be the next step in predicting drug response and personalized endocrine therapies for responsive patients.

Author contribution

This study was designed by AR, GD and PG; the experimental procedures were performed by AR with assistance from SK and BD; microarray, statistical and additional bioinformatics analyses were performed by MA and CE.

Acknowledgements

The authors are grateful to Nard Kubben for optimising and providing us with the protocol for the linear amplification of DNA prior to labelling and chip hybridisation. We are grateful to Prof Scheule, Prof O'Malley and Prof Evans for providing the expression and reporter plasmids we used. This study has been supported by internal funds of the Maastricht University Medical Centre. The authors declare no conflict of interest.

References

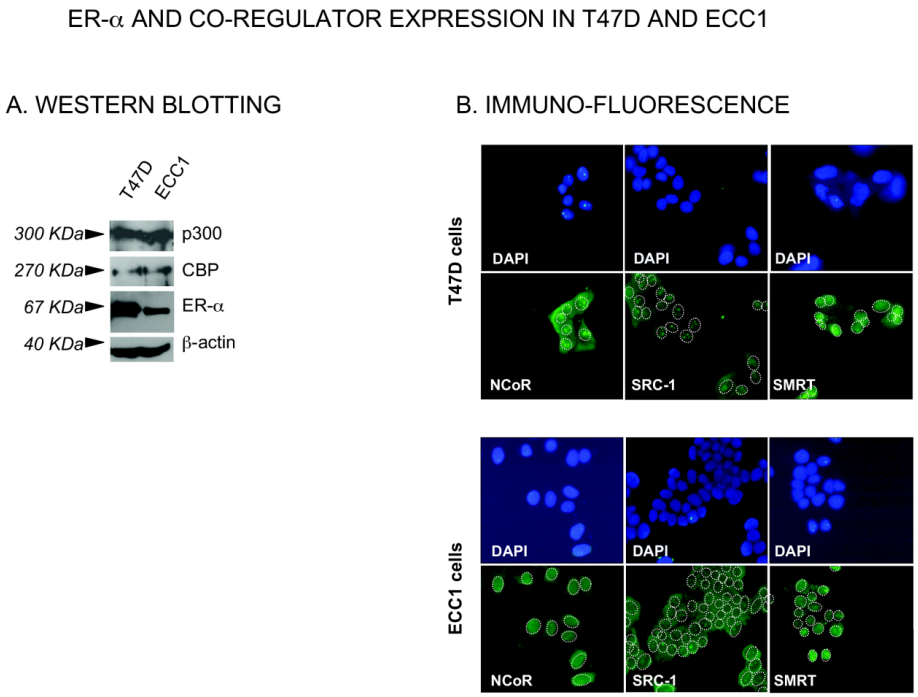
1. Lonard DM, O'Malley B W: Nuclear receptor coregulators: judges, juries, and executioners of cellular regulation. Mol Cell 2007, 27(5):691-700.
2. Riggs BL, Hartmann LC: Selective estrogen-receptor modulators -- mechanisms of action and application to clinical practice. N Engl J Med 2003, 348(7):618-629.
3. Conzen SD: Nuclear Receptors and Breast Cancer. Mol Endocrinol 2008.
4. Shang Y: Molecular mechanisms of oestrogen and SERMs in endometrial carcinogenesis. Nat Rev Cancer 2006, 6(5):360-368.
5. Gielen SC, Burger CW, Kuhne LC, Hanifi-Moghaddam P, Blok LJ: Analysis of estrogen agonism and antagonism of tamoxifen, raloxifene, and ICI182780 in endometrial cancer cells: a putative role for the epidermal growth factor receptor ligand amphiregulin. J Soc Gynecol Investig 2005, 12(7):e55-67.
6. Lonard DM, Lanz RB, O'Malley BW: Nuclear receptor coregulators and human disease. Endocr Rev 2007, 28(5):575-587.
7. Perez-Gracia JL, Carrasco EM: Tamoxifen therapy for ovarian cancer in the adjuvant and advanced settings: systematic review of the literature and implications for future research. Gynecol Oncol 2002, 84(2):201-209.
8. Carroll JS, Brown M: Estrogen receptor target gene: an evolving concept. Mol Endocrinol 2006, 20(8):1707-1714.
9. Shang Y, Brown M: Molecular determinants for the tissue specificity of SERMs. Science 2002, 295(5564):2465-2468.
10. Balmer NN, Richer JK, Spoelstra NS, Torkko KC, Lyle PL, Singh M: Steroid receptor coactivator AIB1 in endometrial carcinoma, hyperplasia and normal endometrium: Correlation with clinicopathologic parameters and biomarkers. Mod Pathol 2006, 19(12):1593-1605.

11. Shang Y, Hu X, DiRenzo J, Lazar MA, Brown M: Cofactor dynamics and sufficiency in estrogen receptor-regulated transcription. Cell 2000, 103(6):843-852.
12. Stossi F, Likhite VS, Katzenellenbogen JA, Katzenellenbogen BS: Estrogen-occupied estrogen receptor represses cyclin G2 gene expression and recruits a repressor complex at the cyclin G2 promoter. J Biol Chem 2006, 281(24):16272-16278.
13. Peterson TJ, Karmakar S, Pace MC, Gao T, Smith CL: The silencing mediator of retinoic acid and thyroid hormone receptor (SMRT) corepressor is required for full estrogen receptor alpha transcriptional activity. Mol Cell Biol 2007, 27(17):5933-5948.
14. Smith CL, Nawaz Z, O'Malley B W: Coactivator and corepressors regulation of the agonist/antagonist activity of the mixed antiestrogen, 4-hydroxytamoxifen. Mol Endocrinol 1997, 11:657-666.
15. Bourdeau V, Deschenes J, Laperriere D, Aid M, White JH, Mader S: Mechanisms of primary and secondary estrogen target gene regulation in breast cancer cells. Nucleic Acids Res 2008, 36(1):76-93.
16. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoute J, Brodsky AS, Keeton EK, Fertuck KC, Hall GF et al: Genome-wide analysis of estrogen receptor binding sites. Nat Genet 2006, 38(11):1289-1297.
17. Hodges LC, Cook JD, Lobenhofer EK, Li L, Bennett L, Bushel PR, Aldaz CM, Afshari CA, Walker CL: Tamoxifen functions as a molecular agonist inducing cell cycle-associated genes in breast cancer cells. Mol Cancer Res 2003, 1(4):300-311.
18. Kwon YS, Garcia-Bassets I, Hutt KR, Cheng CS, Jin M, Liu D, Benner C, Wang D, Ye Z, Bibikova M et al: Sensitive ChIP-DSL technology reveals an extensive estrogen receptor alpha-binding program on human gene promoters. Proc Natl Acad Sci U S A 2007, 104(12):4852-4857.
19. Lin CY, Strom A, Vega VB, Kong SL, Yeo AL, Thomsen JS, Chan WC, Doray B, Bangarusamy DK, Ramasamy A et al: Discovery of estrogen receptor alpha target genes and response elements in breast tumor cells. Genome Biol 2004, 5(9):R66.
20. Lin CY, Vega VB, Thomsen JS, Zhang T, Kong SL, Xie M, Chiu KP, Lipovich L, Barnett DH, Stossi F et al: Whole-genome cartography of estrogen receptor alpha binding sites. PLoS Genet 2007, 3(6):e87.
21. Lin Z, Reierstad S, Huang CC, Bulun SE: Novel estrogen receptor-alpha binding sites and estradiol target genes identified by chromatin immunoprecipitation cloning in breast cancer. Cancer Res 2007, 67(10):5017-5024.
22. Groothuis PG, Dassen HH, Romano A, Punyadeera C: Estrogen and the endometrium: lessons learned from gene expression profiling in rodents and human. Hum Reprod Update 2007, 13(4):405-417.
23. Romano A, Delvoux B, Fischer DC, Groothuis P: The PROGINS polymorphism of the human progesterone receptor diminishes the response to progesterone. J Mol Endocrinol 2007, 38(1-2):331-350.

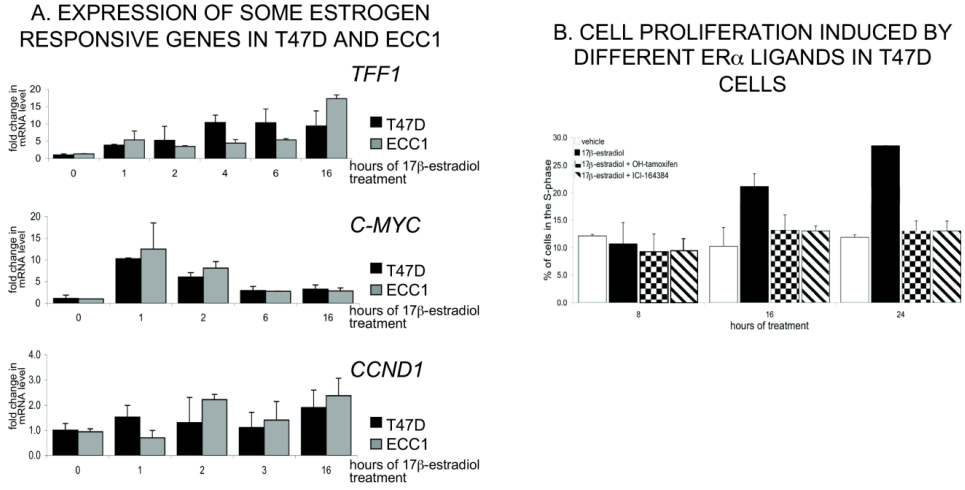
24. Benjamini Y, Hochberg Y: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Roy Statist Soc Ser B 1995, 57:289-300.
25. O'Lone R, Frith MC, Karlsson EK, Hansen U: Genomic targets of nuclear estrogen receptors. Mol Endocrinol 2004, 18(8):1859-1875.
26. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T: MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics 2005, 21(13):2933-2942.
27. Oehler MK, Greschik H, Fischer D-C, Tong X-W, Schuele S, Kieback DG: Somatic Mutations Affecting the Function of the Human Estrogen Receptor alpha(hER-alpha) in Adenomyosis Uteri. Mol Human Reproduction 2004, 10:853-860.
28. Chen JD, Evans RM: A transcriptional co-repressor that interacts with nuclear hormone receptors. Nature 1995, 377(6548):454-457.
29. Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA, Szary AJ, Eeckhoute J, Shao W, Hestermann EV, Geistlinger TR et al: Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. Cell 2005, 122(1):33-43.
30. Smith CL, Onate SA, Tsai MJ, O'Malley BW: CREB binding protein acts synergistically with steroid receptor coactivator-1 to enhance steroid receptor-dependent transcription. Proc Natl Acad Sci U S A 1996, 93(17):8884-8888.
31. Vo N, Goodman RH: CREB-binding protein and p300 in transcriptional regulation. J Biol Chem 2001, 276(17):13505-13508.
32. Cheng ASL, Jin VX, Fan M, Smith LT, Liyanarachchi S, Yan PS, Leu Y-W, Chan MWY, Plass C, Nephew KP et al: Combinatorial analysis of transcription factor partners reveals recruitment of c-MYC to estrogen receptor-alpha responsive promoters. Molecular cell 2006, 21(3):393-404.
33. Jin VX, Leu Y-W, Liyanarachchi S, Sun H, Fan M, Nephew KP, Huang THM, Davuluri RV: Identifying estrogen receptor alpha target genes using integrated computational genomics and chromatin immunoprecipitation microarray. Nucleic acids research 2004, 32(22):6627-6635.
34. Laganière J, Deblois G, Lefebvre C, Bataille AR, Robert F, Giguère V: From the Cover: Location analysis of estrogen receptor alpha target promoters reveals that FOXA1 defines a domain of the estrogen response. Proceedings of the National Academy of Sciences of the United States of America 2005, 102(33):11651-11656.
35. Schultz JR, Petz LN, Nardulli AM: Cell- and ligand specific regulation of promoters containing activator protein-1 and Sp1 sites by estrogen receptors α and β. J Biol Chem 2005, 280:347-354.
36. Klinge CM: Estrogen receptor interaction with estrogen response elements. Nucleic Acids Res 2001, 29(14):2905-2919.
37. Ramsey TL, Risinger KE, Jernigan SC, Mattingly KA, Klinge CM: Estrogen receptor β isoforms exhibit differences in ligand-activated transcriptional activity in an estrogen response element sequence-dependent manner. Endocronology 2004, 145:149-160.

38. Metivier R, Penot G, Carmouche RP, Hubner MR, Reid G, Denger S, Manu D, Brand H, Kos M, Benes V et al: Transcriptional complexes engaged by apo-estrogen receptor-alpha isoforms have divergent outcomes. Embo J 2004, 23(18):3653-3666.
39. Metivier R, Penot G, Hubner MR, Reid G, Brand H, Kos M, Gannon F: Estrogen receptor-alpha directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter. Cell 2003, 115(6):751-763.
40. Yahata T, Shao W, Endoh H, Hur J, Coser KR, Sun H, Ueda Y, Kato S, Isselbacher KJ, Brown M et al: Selective coactivation of estrogen-dependent transcription by CITED1 CBP/p300-binding protein. Genes Dev 2001, 15(19):2598-2612.
41. Kuang SQ, Liao L, Wang S, Medina D, O'Malley BW, Xu J: Mice lacking the amplified in breast cancer 1/steroid receptor coactivator-3 are resistant to chemical carcinogen-induced mammary tumorigenesis. Cancer Res 2005, 65(17):7993-8002.
42. Smith CL, O'Malley BW: Coregulator function: a key to understanding tissue specificity of selective receptor modulators. Endocr Rev 2004, 25(1):45-71.
43. Wang S, Yuan Y, Liao L, Kuang SQ, Tien JC, O'Malley BW, Xu J: Disruption of the SRC-1 gene in mice suppresses breast cancer metastasis without affecting primary tumor formation. Proc Natl Acad Sci U S A 2009, 106(1):151-156.
44. Yu C, York B, Wang S, Feng Q, Xu J, O'Malley BW: An essential function of the SRC-3 coactivator in suppression of cytokine mRNA translation and inflammatory response. Mol Cell 2007, 25(5):765-778.
45. Krum SA, Miranda-Carboni GA, Lupien M, Eeckhoute J, Carroll JS, Brown M: Unique ERalpha cistromes control cell type-specific gene regulation. Mol Endocrinol 2008, 22(11):2393-2406.

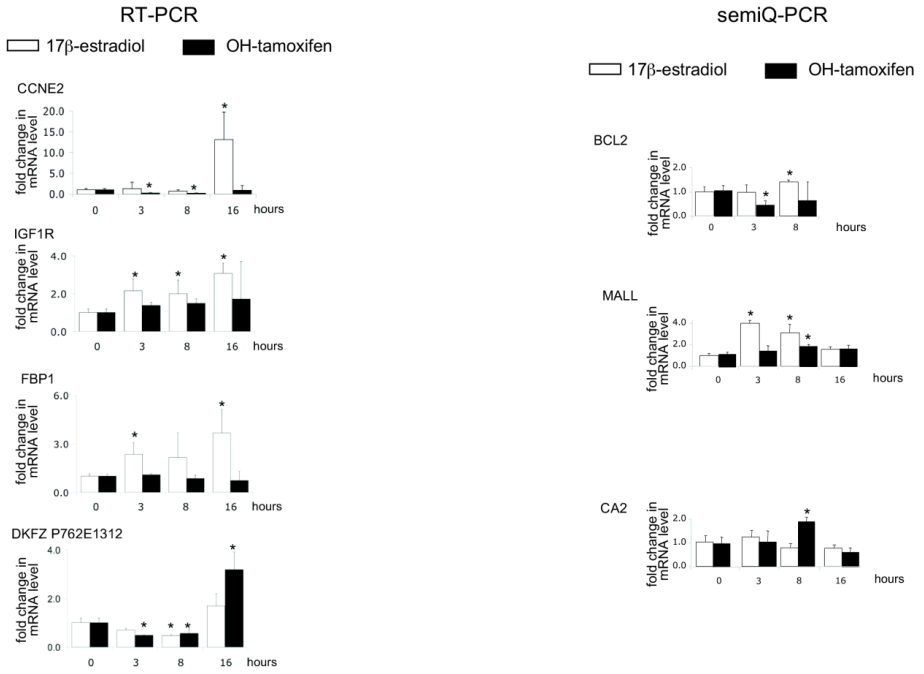
Supplemental figures and tables



Supplemental Figure S-1. *Expression of ER-α and co-regulators in breast cancer T47D and endometrial cancer ECC1 cells. The expression of ER-α, co-activators CBP, p300 and SRC-1 and co-repressors NCoRand SMRT in T47D and ECC1 was assessed by western blot analysis (A) or immunocytochemistry (B). For western blot, anti ER-α monoclonal F10, anti p300 (rabbit, A-22) and anti CBP (rabbit, N-15) antibodies (Santa Cruz Biotechnology, California, USA) were used. Mouse antibody AC-15 (Sigma-Aldrich Chemie BV, Zwijndrecht, The Netherlands) was used to detect β-actin. For visualization of bound antibodies, HRP-conjugated rabbit anti-mouse-antibodies (DAKO, Glostrup, Denmark) and goat-anti-rabbit-antibodies (Pierce, Aalst, Belgium) in conjunction with the super signal-R West-Femto kit (Pierce, Aalst, Belgium) were used. For immunofluorescence, cells were cultured on glass cover slips fixed in buffered formaldehyde (4% paraformaldehyde in PBS), permeabilised with 0.1% Triton-X-100 in PBS and stained with goat polyclonal C-20 against co-activator SRC-1 and sc-1609 against co-repressors NCoR (Santa Cruz Biotechnology, California, USA), followed by anti-goat FITC secondary antibody 705-095-147 (Jackson Immunoresearch/Brunschwig chemie B.V., Amsterdam, The Netherlands); rabbit polyclonal H-300 against co-repressors SMRT (Santa Cruz Biotechnology, California, USA), followed by anti-rabbit FITC F005401 (DAKO, Glostrup, Denmark). Nuclei were stained with 4'-6-diamidino-2-phenylindole (DAPI); nuclei are also indicated by the dashed circles in the antibody staining. Additional technical information are described in the 'Materials and Methods' paragraph or were previously published in Romano et al. (2007) J Mol Endocrinol 38, 331-350.*

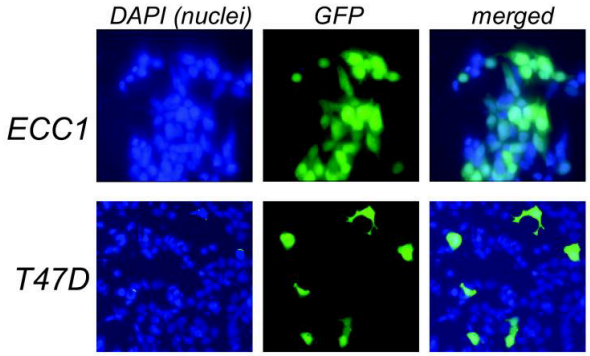


Supplemental Figure S-2. *Response to estrogen stimulation in T47D and ECC1 cells. A. Induction of some known estrogen responsive genes measured by real-time PCR in T47D and ECC1 cells.*
B. FACS analysis of T47D cells treated for the indicated period of time with different ER-α ligands: vehicle-only; 17β-estradiol (1 nM); 17β-estradiol and OH-tamoxifen (1 nM and 1 μM, respectively); 17β-estradiol and ICI-164384 (1 nM and 1 μM, respectively). The percentage of cells in the S-phase under each indicated condition is shown and indicate that T47D cells proliferate with increased rate when treated with 1 nM 17β-estradiol and this effect is impaired by treatment with the SERMs OH-tamoxifen and ICI-134384 (ER-α-antagonists in breast cells). FACS analysis was performed as previously described (Romano et al. 2007, J Mol Endocrinol 38, 331-350) and cell-cycle was analysed by WinMDI and cylchred. Mean ± SD is based on three replicates.



***Supplemental Figure S-3**. mRNA level of a number of ER-α targets identified by ChIP-chip. mRNA level for each indicated gene was assessed by real-time PCR (RT-PCR) or semiquantitative PCR (semiQ-PCR) after induction with 1 nM 17β-estradiol or 1 μM OH-tamoxifen for different periods of time (indicated on the X). Mean ± SD is based on three replicates. Asterisks indicate $p < 0.05$ compared to time point zero.*

Efficiency of transient transfection in ECC1 and T47D cells using the Green-Florescent-Protein (GFP) expression plasmid



***Supplemental Figure S-4**. Transfection efficiency in ECC1 and T47D cells. The transfectability of ECC1 and T47D cells was confirmed by observing cells expressing the green fluorescence protein (GFP) after transfection with a GFP reporter plasmid.*

Supplemental Table I. Number of ER-α binding sites per chromosome.

Chromosome	*number of sites*	*Chromosome*	*number of sites*
1	79	13	17
2	51	14	33
3	45	15	31
4	36	16	27
5	36	17	54
6	61	18	12
7	41	19	41
8	35	20	31
9	43	21	12
10	25	22	21
11	83	X	34
12	56		

Additional supplemental tables S-I, S-II, S-III and S-IV available from
http://www.sciencedirect.com/science/article/pii/S0303720709004122

Chapter 7

Hypoxia induces bivalent chromatin domains by specific gain of H3K27me3

Peggy Prickaerts*[1], Michiel Adriaens*[2], Twan van den Beucken[3,4,5,6], Vivian EH Dahlmans[1], Michelle Chan-Seng-Yue[3], Bradly G Wouters*[#3,4,5,6], Jan Willem Voncken*[#1]

** Equal contribution*

[#] To whom correspondence may be addressed

[1] Department of Molecular Genetics, Maastricht University, Maastricht, The Netherlands

[2] Department of Bioinformatics – BiGCaT, Maastricht University, Maastricht, The Netherlands

[3] Ontario Institute for Cancer Research, Toronto, ON, Canada

[4] Ontario Cancer Institute and Campbell Family Institute for Cancer Research, Princess Margaret Hospital, University Health Network, Toronto, Canada

[5] Maastricht Radiation Oncology (MaastRO) Lab, Maastricht University, Maastricht, The Netherlands

[6] Departments of Radiation Oncology and Medical Biophysics, University of Toronto, Toronto, ON, Canada

Keywords: *Deep sequencing, hypoxia, reoxygenation, epigenetics, bivalency, H3K4me3, H3K27me3.*

Publication: Manuscript in preparation.

Abstract

Trimethylation at histone H3 lysine 4 (H3K4me3) and lysine 27 (H3K27me3) has been linked to gene activity and repression, respectively. Distinctive H3K4me3 and H3K27me3-enrichment at specific target-genes in embryonic stem cells and more committed cell lineages reflects spatio-temporal epigenetic control over developmental processes and demonstrates that genomic distribution of both epigenetic marks is subject to dynamic change. How H3K4me3 and H3K27me3 respond to changes in the microenvironment is relatively unknown. Based on the biochemical dependency of the Jumonji-class histone demethylases, we hypothesized that genome-wide histone trimethylation enrichment will be dynamically affected by changes in cell oxygenation. We have determined the relationship between epigenomic and transcriptomic reprogramming in a model for fluctuating oxygen tension within the tumor micro-environment. To this end, we have combined chromatin-immunoprecipitation and deep-sequencing analysis of H3K4me3 and H3K27me3-enrichment with expression array data of MCF7 breast cancer cells subjected to changes in oxygen tension (*i.e*. acute hypoxia, chronic hypoxia and reoxygenation. We observed a rapid global increase of both H3K4me3 and H3K27me3-marks at specific sites throughout the genome, which was reversed upon reoxygenation. Normoxic H3K4me3-profiles at marked genes were only marginally affected by hypoxia.. Acquisition or loss of H3K4me3 at target genes correlated with increased or reduced gene expression, respectively. In sharp contrast, *de novo* genic H3K27me3-marking was found to accumulate around transcription start sites (TSS) during hypoxia, and was transitory in nature and did not correlate with transcriptional repression. Thus, under conditions of oxygen deprivation, H3K4me3-occupation was identified as the most important epigenetic marker of transcriptional regulation. As many TSS loci were already H3K4me3-marked, acquisition of H3K27me3 resulted in increased bivalent marking. Hypoxia-induced bivalency revealed substantial overlap with embryonal stem cell-associated bivalency and was retained at numerous loci upon reoxygenation. This suggested strict control over histone (de)methylation at these sites. Our data show for the first time that oxygen availability dynamically regulates the epigenetic state of the genome. The possible repercussions of hypoxia-induced bivalency in the context of acquisition of stem cell-like epigenomic marking and tumor plasticity is discussed.

Introduction

Cancer cells in solid tumors are often exposed to fluctuating oxygen tension resulting from inadequate blood supply due to poorly developed vasculature [1]. Transcriptional changes in hypoxic cancer cells are influenced through several well understood hypoxia response pathways, including stabilization ad activation of hypoxia-inducible factor 1a (HIF-1a) [2, 3]. Transcriptional modulation of genes involved in glycolysis, angiogenesis, pH homeostasis and apoptosis (i.e. anti-apoptotic genes) enable cancer cells to survive and adapt to the hypoxic environment. Repeated oxygen deprivation and reoxygenation also has been hypothesized to promote tumor stem cell properties, metastasis, and patient prognosis. The phenotypic changes induced by adaptive responses to oxygen deprivation, in combination with other

mutational changes in cancer, severely decrease the effectiveness of both ionizing radiation and chemotherapy [4, 5].

Epigenetic regulation of gene expression is coordinated at the level of DNA methylation, covalent histone modifications and expression of non-coding RNAs. Their concerted action affects transcriptional regulation by influencing both access and binding of regulatory factors to chromatin as a result of changed chromatin-compaction, repositioning of nucleosomes and/or recruitment of regulatory factors (i.e. by acting as a scaffold). These principles also apply to other DN-templated processes such as replication and repair [6, 7]. Recent data has demonstrated that epigenetic regulation can also mediate adaptation to changes in the micro-environment and this feature constitutes a major underlying mechanism in development, maintenance of cellular diversity, phenotypic plasticity and homeostasis [8].

The epigenetic status of the genome, including all chemical modifications of DNA and histone proteins, is often referred to as the *epigenome* [9]. Histone methylation, acetylation and phosphorylation are well-documented covalent chemical modifications that occur on N-terminal tails of core histones that constitute the nucleosomal units in chromatin [7]. Histone acetylation is firmly connected to gene activation. In contrast, histone methylation is associated with active gene transcription as well as gene repression: H3 lysine 9 (H3K9), H3K27 and H4K20 trimethylation (me3) are generally associated with gene silencing, whereas H3K4, H3K36 and H3K79 me3 are linked to transcriptional activation [6, 7].

H3K27me3 and H3K4me3 states have been intensely studied in the context of development in relation to transcriptional regulation [10-13]. H3K27 is trimethylated by Polycomb Repressive Complex 2 (PRC2), which comprises the histone methyl transferase EZH2 [14]. The reverse process (i.e. H3K27-demethylation) is accomplished by the histone demethylases UTX and JMJD3 [15]. H3K4me3 marks are installed by MLL-proteins that belong to the Trithorax Group (TrxG) of epigenetic modifiers, which functionally counteract PRCs [16, 17]. JARID1A-D catalyse H3K4 demethylation [18]. Although in terminally differentiated cells most H3K4me3 and H3K27me3 marks appear mutually exclusive, bivalent marking at key developmental control genes is observed in embryonic stem cells (ESC) [10-13]. Epigenomic comparison between ESC and committed and/or fully differentiated cell types has shown that bivalent marking is resolved to monovalency (or loss of both marks) at some stage during lineage commitment and differentiation. Eventually this results in lineage-specific H3K27me3 and H3K4me3-profiles, which are stably transmitted during cell division [13, 19]. As such, PRCs and TrXG play a fundamental role in the establishment and maintenance of lineage-specific gene expression [16, 17].

Current knowledge on the distribution dynamics of H3K27me3 and H3K4me3 beyond embryonic development is relatively limited. We therefore studied the dynamics of histone methylation in the context of microenvironmental change. As a relevant physiological model, we used cellular adaptation to hypoxia and subsequent reoxygenation. Acute and chronic hypoxia are known to induce major transcriptomic changes [20]. Furthermore, hypoxia was reported to induce global changes in histone methylation [21]. However, a systematic profiling of H3K27me3 and H3K4me3 in relation to hypoxia induced gene expression is lacking. Relevantly, removal of histone trimethylation states is directly coupled to cell

oxygenation: the Jumonji C-terminal domain histone demethylases (JHDM) use Fe2+, α-ketoglutarate and oxygen as co-substrates to remove all methylation states by hydroxylation [22-26]; as such the regulation of of α-ketoglutarate-dependent demethylases under low oxygen is expected to be very similar, if not identical to that of the prolylhydoxylase HIF-PH which targets hypoxia-inducible factor 1a (HIF1a) [27, 28]. We therefore hypothesized that oxygen affects the activity of JHDMs and will increase global repressive and/or activating histone trimethylation-marking. Since HIF1a targets transcription of a subset of responsive genes, it is expected that constitutive transcriptional activity is required at corresponding genic areas and that these need to be exempted from silencing.

To study the relation between epigenomic and transcriptomic changes, we employed chromatin-immunoprecipitation (ChIP) followed by deep-sequencing (ChIP-seq) [29, 30] and combined this with expression array analysis. We charted the distribution of both histone marks as a function of time cultured under low oxygen and correlated their distribution profiles to our gene expression data. Of relevance, we also included reoxygenation in our measurements, as cancer cells are subject to constantly changing oxygen tension in the tumor microenvironment. As H3K27me3 is known to cover large chromosomal regions instead of sharp defined peaks [31, 32], we also developed a standardized protocol to define and summarize H3K27me3-enrichment using deep-sequencing analysis. A detailed description of the protocol and the scripts used are published elsewhere [33].

Methods

Cell Culture, expression vectors and viral infections

MCF7 (human mammary adenocarcinoma) and DU145 (human prostate carcinoma) cells (ATCC) were cultured at 37°C, 5% CO_2, 100% humidity in Dulbecco's Modified Eagle Medium**:**Nutrient Mixture F-12 (DMEM/F12 1:1; MCF7) and DMEM McCoy's 5A medium (DU145). The culture medium was supplemented with 10% fetal calf serum (FCS), 200 mM L-glutamine and antibiotics For hypoxic exposure, cells were transferred to a MACS VA500 microaerophilic workstation (Don Whitley Scientific, Shipley, UK) for the indicated duration. The atmosphere in the chamber consisted of <0.02% O_2, 5% H_2, 5% CO_2, and 74% N_2. For reoxygenation, cells were transferred back to the regular tissue culture inhibitor containing ambient oxygen levels (21%). RNA interfering HIF1a sequences were obtained from Sigma (clone TRCN0000010819). Plasmids and lentiviral work Knock-down of HIF1alpha was achieved using lentiviral shRNA constructs. Lentiviral particles were generated by co-transfection of 293T cells with packaging plasmids pCMVdR8.74psPAX2 and pMD2.G together with shRNA vector pLKO.1. Virus supernatant was harvested 48 and 72 hrs post transfection. MCF7 cells were transduced with lentiviral supernatant in the presence of 8 microg/ml polybrene. Infected cells were selected for 2 days in 2 microg/ml puromycin containing media. Efficiency of plasmid constructs was verified by immunoblotting.

Protein isolation and Western blot analysis

Cells were grown to around 70% confluency before they were transferred to the hypoxic chamber. Protein extraction was carried out using RIPA buffer supplemented with protease and phosphatase inhibitors. Lysates were further processed and protein concentrations were determined. Immunoblotting (IB) was performed as described previously, using antibodies raised against HIF1a ( ), H3K4me3 (Ab8580; Abcam, Cambridge, UK), H3K27me3 (07-449; Upstate Biotechnology/Millipore, Waltham, MA, USA), H3K9/K14ac (Abcam, Cambridge, UK), H3 (ab1791; Abcam, Cambridge, UK), b-Actin (C4, 69100, MP Biomedicals, Solon, OH, USA),. For more details, see *Supplementary Methods*.

Chromatin immunoprecipitation (ChIP) assays

MCF7 cells were transferred to hypoxic culturing conditions for the indicated durations and immediately fixed to avoid reoxygenation. Cells were disrupted by sonication, yielding genomic DNA fragments ranging from 200-1000 bp, with a bulk size of 200-500bp. For each immunoprecipitation 10-20 million cells were used. 1% of the cell suspension was kept aside as input DNA to use as a reference. ChIPs were performed and analyzed as described previously with minor adjustments [31], see supplementary materials and methods. Antibodies used include: H3K4me3 (Ab8580; Abcam), H3K27me3 (07-449; Upstate), CBX8 (LAST; courtesy Klaus Hansen, Copenhagen, Denmark) and HA as a negative control (sc-805; Santa Cruz Biotechnology, Santa Cruz, CA, USA). The immunoprecipitated DNA was checked for enrichment using real-time PCR and quantified by fluorescence detection using Quant-iT™ Picogreen® dsDNA Reagent (Molecular Probers/Invitrogen, Eugene, OR, USA) before deep sequencing was applied.

Deep sequencing

Input and ChIP samples were further processed and sequenced at the Ontario Institute for Cancer Research using the Illumina next generation sequencing platform. Processing involved size selection to enrich for mononucleosome-sized fragments, linker annealing and PCR amplification. Each resulting library was subsequently loaded onto individual lanes of a flow cell and sequenced using the 36 bp paired-end protocol on the Illumina Genome Analyser IIx (GAIIx). In order to obtain sufficient sequencing depth additional lanes were sequenced if necessary. All data obtained from each individual sample was pooled. Data sets will be made publically available.

Genome alignment, normalization, background correction; identification of enriched regions

Image processing and base calling was performed using Illumina software tools provided by the manufacturer). Subsequent paired-end genome alignment was performed using Novoalign with Human Genome 18 (HG18) used as a reference genome. Only uniquely aligned reads were used for further analysis. To remove PCR artifacts all data were collapsed prior to peak calling. H3K27me3 data sets were normalized based on identification of regions with stable H3K27me3-enrichment between all

samples analyzed; the cumulative area under the curve for all peaks within these regions was scaled relative to the smallest value among the samples; the normalization strategy will be outlined in detail elsewhere [33]. After normalization, a single cut-off value for all samples was set at the enrichment level (peak height; estimated by input sample data) above which H3K27me3 signal correlates with a known H3K27me3-binding protein under normoxic conditions (t=0; data not shown). Signal intensities below this cut-off were not considered biologically relevant. To identify enriched regions in the ChIP samples relative to the input control, the peak caller Findpeaks (version 4.0) was used. For H3K4me3 the default settings were used, whereas for H3K27me3 the settings were adjusted in order to detect blanketing enrichment next to sharply defined peaks [33].

Microarray

RNA for microarray application was isolated using RNeasy mini kit (Qiagen, Hilden, Germany) according to manufacturer's protocol. Isolations were performed in triplicate. Total RNA samples were analysed using the Affymetrics expression array platform (Affymetrix Gene Chip 1.0 ST). After scanning, data preprocessing and data analysis were done with R (http://www.R-project.org; version 2.12) using the Bioconductor (http://www.bioconductor.org; version 2.7). Data were background-corrected and normalized using gcRMA [34]. Microarray data will be made available at Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/).

Integration with gene expression; gene ontology analysis

To enable integration of expression data with our enrichment data, all Affymetrix prob IDs were converted into ENSEMBL gene identifiers. The expression level of an individual gene is defined as the average of all probe sets representing this gene on the array. Genes were considered expressed if expression exceeds 100 for at least one independent time-point. Genes were called regulated if genes are expressed (>100 for 1 or more time-points) and the fold change between 2 independent time-points is ≥ 2. All genes which were not represented on the micro-array were not included for further analysis. For the identification of enriched genes, we defined a gene as the region between its 5' (most upstream TSS) and 3' (last exon) end plus 5 kb regulatory regions up and downstream respectively. A gene was called marked when there was a peak present within this region as determined by the enrichment finding procedure. Gene Ontology enrichment analysis was performed using topGO [35, 36]. Comparative GO analysis with embryonal stem cell bivalent markers was based on published data [12].

Data visualization

For data visualization several plots were created in R, including histograms, gene tracks, genome plots, pie charts, TSS plots, boxplots, scatter plots and pie-charts. For the TSS plots all genes were considered in the same orientation (from 5' to 3' end), and the average signal intensity was depicted for a region defined in number of base pairs surrounding the TSS, as indicated. Box-plots show the 25^{th} and 75^{th}

quartile (as indicated by the box), the median (indicated as a line within the box), and the whiskers indicate the 5% and 95% percentile respectively. Notches in box-plots indicate confidence intervals (5-95%) of the median.

Data summarization

To study different enrichment profiles for H3K27me3, each gene was assigned to the Promoter, the TSS or the Broad class in analogy to published analysis [37]. To properly classify all genes, each gene was first divided into three regions: the promoter region (-3000/-100 base pairs (bp) in relation to the TSS), the TSS region (-100/1000 bp), the broad region (+1000 bp to the last exon). Genes shorter than 4000 bp were excluded from the analysis, as they were too small to reliably assign them to either profile. Genes were allocated to the different classes based on which of the three regions contained the largest amount of signal scaled to the size of each region, provided it contains a peak which showed at least 25% more enrichment compared to any other peak within the gene (Promoter and TSS class), or it contains a enrichment above the average enrichment at more than 35% of the region (Broad class). Genes allocated to either region which did not meet above criteria, were not considered for profile analysis.

Results

Reversible oxygen-dependent global changes of H3K4me3 and H3K27me3 levels

To determine whether histone trimethyl-states are dependent on the availability of molecular oxygen we exposed MCF7 breast cancer cells to severe hypoxia (<0.02%) and measured global changes in H3K4me3 and H3K27me3 levels by immunoblotting at 8 and 24 hours of hypoxia compared to normoxic cells (t=0). Cancer cells are subject to repetitive fluctuation of oxygen availability *in situ* (*i.e.* hypoxia and reoxygenation); reoxygenation may affect cells through specific stress responses that that involve epigenetic change. For this reason we also determined H3K4me3 and H3K27me3 levels in response to reoxygenation. Trimethylation of H3K4 (1.7 fold) and H3K27 (2.2 fold) was increased after 8 hours culturing at low oxygen tension; the elevated trimethyl-state was sustained up until 24 hours of culturing under hypoxic conditions (**Figure 1A**). We observed this in multiple cell lines, suggesting that the epigenomic remodeling was dictated by oxygen-status and occurred independent of cellular context (**Figure S1A**). Importantly these initial observations are consistent with our original hypothesis. Conversely, restoration of oxygenation induced an initial sharp decline of global histone H3K4 and K27-trimethylation that transiently dropped below levels in normoxic cells and returned to baseline at approximately 12-24 hours after reoxygenation, depending on the cell type used; concomitantly, H3K9/K14-acetlyation (H3K9/K14a) increased in response to oxygen stress. (**Figure 1A, Figure S1A**).

To establish that demethylation activity is reduced because of loss of oxygen, and not by changes in protein levels of the responsible molecular machineries, histone lysine demethylase (HKDM) and histone lysine methyl transferase (HKMT) mRNA levels were measured. Expression of a number of relevant HKMDs including JARID2, JARID3 and JMJD3, show significantly increased expression in hypoxic cells

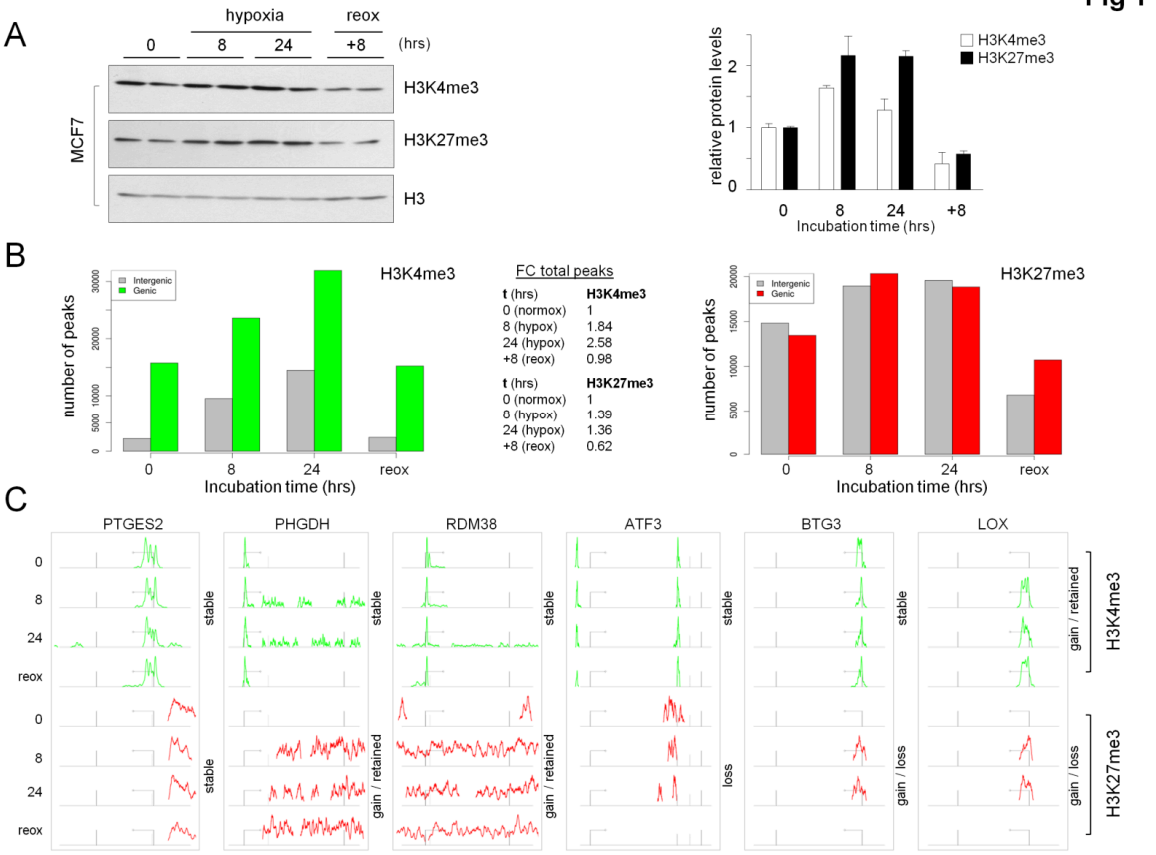


Figure 1. *Reversible oxygen-dependent global changes of H3K4me3 and H3K27me3-levels.* ***A****; Immuno-blot detection (IB) of epigenetic changes (H3K4me3 and H3K27me3) in MCF7 cells; right panel: quantification.* ***B****; H3K4me3 (left panel) and H3K27me3 (right panel) the number of peaks detected significantly above background level (p < 0.05) at different time points under hypoxic conditions (t=8, t=24 hrs) and after reoxygenation (reox; t=+8 hrs).* ***C****, Representative gene tracks of loci displaying (left to right): stable H3K27me3 and H3K4me3-marking, H3K27me3-gain/retained (at reox) and H3K4me3 gain/loss, H3K27me3-gain/retained, loss of K27me3 (stable H3K4me3), H3K27me3-gain and loss (reox), gain of both marks (H3K4me3 retained; reox). Symbol indicates transcription-direction.*

(**Figure S1B**). In contrast, the H3K27me3-HMT EZH2 as well as numerous confirmed and putative H3K4me3 HMTs declined in response to oxygen deprivation or remained unaltered (**Figure S1C**). Hence, changes in levels of responsible writers or erasers could not explain the global increase in both activating and inactivating marks; instead the data is consistent with changes in enzymatic activity of KDM. Moreover, the sudden drop in H3K4me3 and H3K27me3 levels following reoxygenation is in agreement with activation of accumulated HKDM protein by molecular oxygen.

Previous studies have shown that expression of HKDM is controlled by HIF1a [38-40]. To determine whether the effects of altered oxygenation on global trimethylation levels were dependent on HIF1a, HIF1a-depleted MCF7 cells were exposed to hypoxia and analyzed for global changes in H3K4me3 (**Figure S1D**). No obvious differences with respect to hypoxia-induced increased histone trimethylation were observed between control and *shHIF1a* cells, indicating that the global epigenetic changes occurred in a HIF1a-independent fashion (data not shown). Combined, these data suggest that hypoxic stress induces a reversible increase in histone H3 trimethylation which is HIF1a-independent.

We next aimed to determine whether changes in oxygen regulate epigenetic states at specific genes. To this end ChIP-seq analysis was performed on H3K4me3 and H3K27me3-enriched sequences in normoxic (t=0), hypoxic (t=8 and 24 hrs) and reoxygenated cells (t=+8 hrs); the latter condition was considered as a transition point to restoration of the original epigenomic state under ambient conditions (21% oxygen). In parallel, RNA samples were collected for analysis at all four time points. ChIP-seq analysis confirmed enhanced global trimethylation of H3K4 and H3K27 in response to hypoxia, consistent with the immunoblotting findings: the total amount of H3K4me3 and H3K27me3 peaks had increased ±2.6 and 1.4 fold, respectively, at 24 hours hypoxia (**Figure 1B,** *cf.* genome plots **Figure S2A,B)**. Conversely, reoxygenated samples revealed a steep drop in the global number of trimethylation peaks: both the number of H3K4me3 and H3K27me3 peaks fell ±2.5 fold at 8 hours reoxygenation (t=+8) to 0.98 and 0.62 fold their original normoxic values, respectively (**Figure 1B**, *cf.* genome plots **Figure S2A,B**). Representative genome tracks illustrate examples of individual genes that displayed loss, gain or unchanged H3K4me3 and/or H3K27me3 levels in response to altered oxygenation (**Figure 1C**). Consistent with the global decline of H3K4me3 and H3K27me3 (*cf.*,**Figure 1A**) enhanced trimethylation was lost upon reoxygenation at most of these loci. Reciprocal correlation analysis revealed a substantial overlap between enriched loci (for both H3 marks) under normoxic *vs.* hypoxic conditions (corr.coeff. at t=0 *vs* t=24 hrs: 0.45 H3K4me3, 0.53 H3K27me3) and *vice versa* (corr.coeff. at t=24 *vs* t=0 hrs: 0.77 H3K4me3, 0.60 H3K27me3). This data indicated that pre-existent (normoxic) H3K4me3 and K27me3-marking was generally retained under hypoxic conditions and that the increased enrichment was derived from hypoxia-induced *de novo* trimethylation. These findings validated the reliability of the comparative histone me3-marking analysis. H3K4me3-enrichment (*i.e.* sequences associated with peaks) returned to the normal situation normal (corr.coeff. at normoxia vs. reoxygenation and vice versa: 0.82). In contrast H3K27me3-marks show poor correlation between t=0 and t=+8 hrs (corr.coeff: 0.19 norm vs. reox and 0.11 reox vs. norm) indicating that normoxic H3K27me3-distribution has not been restored at 8 hours

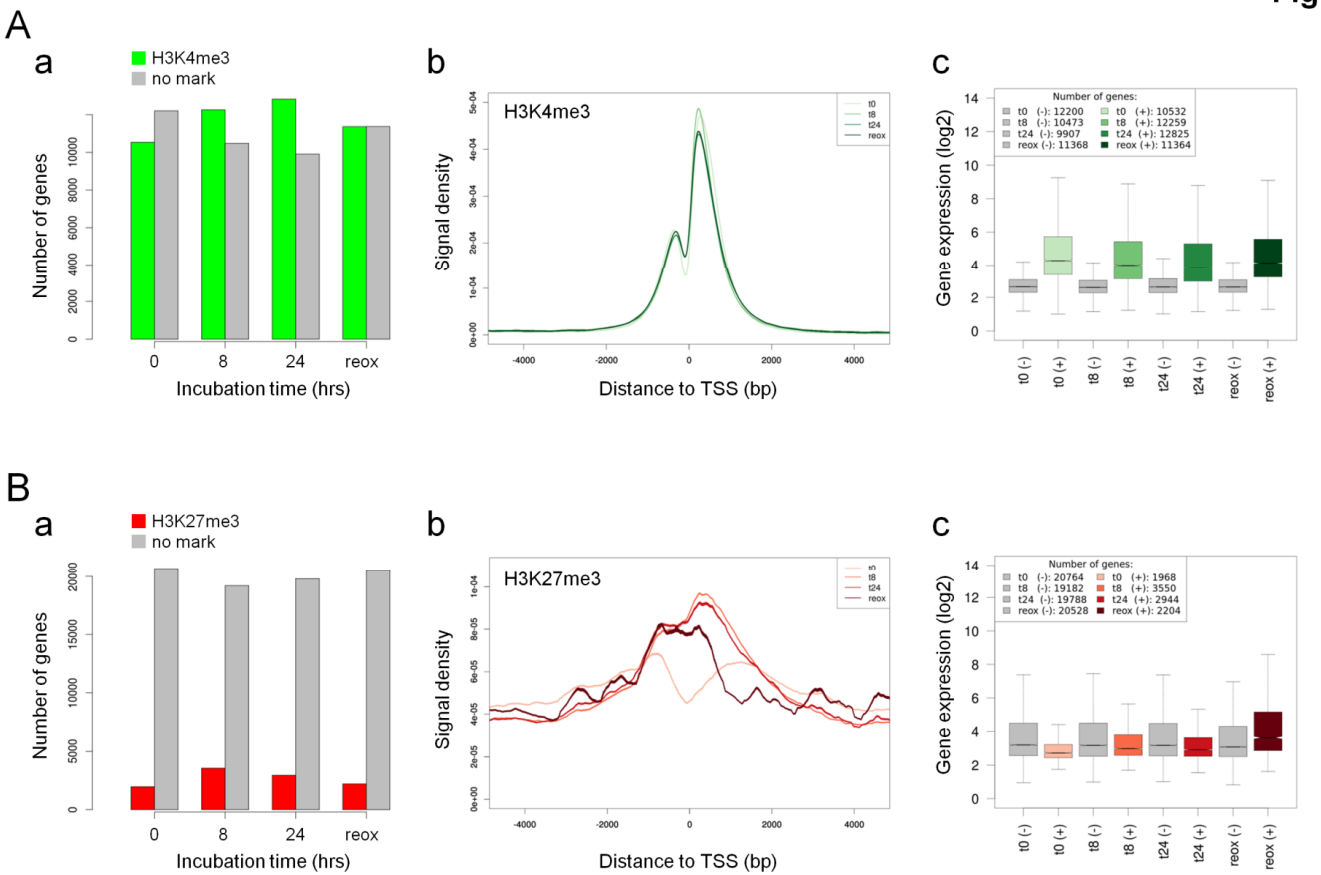


Figure 2. *Preferred gain of histone methylation at genic regions.* ***A****;* ***a****) number of H3K4me3-associated genes;* ***b****) H3K4me3-signal density distribution proximal to the TSS in relation to oxygen deprivation and reoxygenation;* ***c****) median gene expression in relation to H3K4me3-marking.* ***B****;* ***a****) number of H3K27me3-associated genes;* ***b****) H3K27me3-distribution proximal to the TSS in relation to oxygen deprivation and reoxygenation;* ***c****) box-plots presenting median gene expression in relation to H3K27me3-marking; all data are presented at time points indicated (t=8 or 24 hours hypoxia; reoxygenation at t=+8 hours).*

after reoxygenation. Although a number of densly trimethylated genomic regions were relatively resistant to demethylation at reoxygenation, combined, these findings showed that hypoxia causes an overall reversible global increase of H3K4me3 and H3K27me3 and establishes that oxygen is required for constitutive activity of HKDM and maitenance of the basal epigenetic state in a large number of genomic regions.

Preferred gain of histone methylation at genic regions

The above findings provided us with a basis for comparative epigenomic and transcriptomic analysis. To relate epigenomic changes to transcriptional states, we first differentiated enriched sequences into genic (*i.e.* "genes" +/- 5000 bp) and intergenic regions. Under control conditions, trimethylation was highly associated to genic regions: 88% of all H3K4me3 peaks located to genes compared to 47% for H3K27me3 peaks (**Figure 1B**). In normoxic cells, 46% of all genes carried H3K4me3-marks, whereas significantly less genes (9%) were H3K27me3-enriched (**Figure 2A,B/a**). Intergenic H3K4me3-association increased from ±10% to ±30%; the relative genic/intergenic H3K27me3-distribution did not change as a result of oxygen deprivation (±50%; **Figure 1B**). Although overall H3K4me3-peaks associated to genic regions decreased from 88% to 69% in low oxygen conditions (**Figure 1B)**, the number of genes H3K4me3-associated genes increased from 46% to 56% under at 24 hours hypoxia (**Figure S2Aa**). In contrast to the rather stable relative number of H3K27me3 peaks associated with genic regions (48% at normoxia vs 49% at 24 hours hypoxia; **Figure 1B**), the percentage of genes associated with H3K27me3 nearly doubled at 16% after 8 hours of oxygen deprivation (**Figure 2Ba**). Thus, although the percentage of marked genes increased for both methyl marks (**Figure 2A,B/a**), the relative increase of H3K4me3-peaks was higher in intergenic regions, whereas relative gain H3K27me3 occurred mostly in genic regions (**Figure 1B; Figure S3A,B**). Alignment of all trimethylation-associated sequence tags to genic or intergenic regions, showed that the majority (>98%) of trimethyl-marks was directed towards genic regions, irrespective of oxygen tension (**Figure S3C,D**). The combined above data confirms that genomic decoration with H3K4me3 and H3K27me3 relates to gene regulation and indicates that H3 trimethylation does not occur in a random fashion but is strongly directed towards genic regions.

Upon reoxygenation the overall number of trimethyl-marked genes reverted to that measured in normoxic cells (15752 genes, normox; 31919 genes, 24 hrs hypox *vs.* 15236 genes, reox, respectively) (**Figure 1B, Figure 2A-B/a**). The intergenic:genic ratio of trimethyl-mark distribution of H3K4me3 was nearly fully restored to that initially observed at ambient oxygen levels (0.14, t=0; 0.45, t=24 *vs.* 0.16, reox); in contrast, a relative genic-enrichment H3K27me3 was apparent at the 8 hours reoxygenation time point (49% *vs.* 61%) (**Figure 1B**). This data is consistent with the poor H3K27me3-peak correlation between normoxic and reoxygenated samples (noted above) and shows that hypoxia-induced H3K27me3-marking at genic regions is relatively resistant to restored HKDM activity.

Epigenetic profiles correlate with transcriptional state

To profile the oxygen tension-induced enrichment of H3K4me3 and H3K27me3 at genic regions in more detail, enrichment data were visualized using TSS-centered plots. Under control conditions, H3K4me3 was prominently enriched around the TSS with a distinctive signal-depletion directly over the TSS; this finding is consistent with earlier reports showing reduced nucleosome presence at this site (**Figure 2Ab**) [41]. In sharp contrast, H3K27me3-marking at normoxia coincided mainly within gene body-enrichment, consistent with the reported "blanketing" enrichment of H3K27me3 (**Figure 2Bb**) [31-33]. The H3K4me3-enrichment profile did not significantly change as a result of hypoxia or reoxygenation (**Figure 2Ab**), whereas H3K27me3-enrichment was selectively enhanced over and around the TSS during oxygen deprivation (**Figure 2Bb**). Taken together, this data confirmed the genic bias of histone H3 trimethylation at K4 and K27 and revealed a pronounced TSS-directed hypoxia-induced increase of H3K27me3

To correlate H3K4 and H3K27-trimethylation to gene transcription, the ChIP-seq data were compared to expression data of corresponding genes. Expression array analysis revealed that a significant number of transcripts was down-regulated in response to upon hypoxic exposure (**Figure S4A**). Gene ontology (GO) classification confirmed regulation of expected processes in response to hypoxia (**Figure S4C**), consistent with published data [20]. H3K4me3-marked genes were significantly higher expressed at all individual time points (**Figure 2Ac**), whereas H3K27me3-marked genes were transcribed at substantially lower levels as compared to non-marked genes, except for the reoxygenation time point (**Figure 2Bc**). H3K4me3-enrichment showed a positive correlation with gene expression level at all time-points analyzed (data not shown). There was no apparent correlation between relative H3K27me3-enrichment and gene expression/repression at any time point suggesting that despite the overall increased H3K27me3-enrichment during oxygen deprivation, H3K27me3-enrichment was not repressive *per se*.

Reoxygenation only partially restored the hypoxia-induced epigenetic marking and expression levels, indicating that although many cellular processes are affected by reoxygenation, cells are at a transition phase at the 8 hours reoxygenation time point (**Figure S4B,D**). Consistent with the earlier noted sustained H3K27me3-marking during reoxygenation, the TSS-associated H3K27me3-enrichment profile was not immediately reversed at t=+8 hours (reox), confirming that specifically the K27me3 signal-distribution at the TSS was relatively resistant to the effects of reoxygenation compared to overall H3K27me3-marking. The loss of genic H3K4me3-marking at reoxygenation, despite increase overall expression suggested functional uncoupling of epigenetic marking and gene expression in response to acute reoxygenation stress.

Hypoxia induces bivalency

As 46.3% of genes (22732) was already H3K4me3-enriched at normoxia (**Figure 2Ab**) and H3K27me3-enrichment specifically increased around the TSS (**Figure 2Bb**), the hypoxia-induced increase in trimethylation was likely to increase the frequency of co-occurrence of both trimethyl marks. Therefore, we selectively examined genes that showed double marking during hypoxia and calculated the *Pearson*

correlation coefficient (*pcc*) at different time points to determine co-occurrence of trimethyl marks. *Pearson* correlation analysis in control samples (normoxia) indicated a negative correlation between 800 bp up- and downstream of the TSS, corresponding to H3K4me3 enrichment and H3K27me3 depletion around the TSS (**Figure S5A**, *cf.* **Figure 2Ab**). Regions more distant (up- or downstream) from the TSS and the TSS itself showed a positive *pcc*, which correlated with low presence of both marks (TSS) and absence of H3K4me3 (distant). H3K4me3-monovalency around the TSS at t=0 was lost during hypoxia, consistent with the increased H3K27me3-enrichment around the TSS in response to hypoxia (**Figure S5A**, *cf.* **Figure 2Bb**). Restoration of ambient oxygen tension correlated with a sharp *pcc* decline at the TSS, correlating with a relative loss of one of two me3-marks at the TSS itself (**Figure S5A**, *cf.* **Figure 2Ab,Bb**).

To study the co-occurrence of H3K4me3 and H3K27me3 in further detail, all genes which were enriched for both trimethyl marks were plotted for each time point separately. We detected an almost 3 fold (821 genes, t=0 to 2200, t=24) increase in the number of double-marked genes as a result of hypoxia. Remarkably, double-marking was maintained at more than 70% of genic loci upon reoxygenation (t=24 hrs *vs.* reox; **Figure S5B**). This data demonstrated that hypoxia increases co-occurence of H3K4 and H3K27-trimethylation marks, and suggested that this epigenetic state was selectively maintained upon reoxygenation despite global loss in trimethylation.

Transcription at bivalent genes is primarily determined by H3K4me3

Expression of double-marked genes was marginally affected by hypoxia: 16% of double-marked genes were actively expressed in control samples *vs.* 15% and 13% at 8 and 24 hours hypoxia, respectively. To determine the consequences of dynamic epigenetic change, including the co-occurrence of H3K4me3 and H3K27me3, for gene-expression, genes were grouped based on their enrichment profile at normoxia (*i.e.* non-marked, K4me3 only, K27 only, double marked); genes were then compared for changes in average expression levels in relation to changes in epigenetic marking. 90% of genes which were H3K4me3-marked at normoxia showed persistent H3K4me3-marking well into hypoxia (t=24 hrs) and remained transcriptionally active (**Figure 3A**). Notably, loci at which H3K4me3 was lost (single H3K27+ or non-marked at t=24 hrs) were transcriptionally silenced. In sharp contrast, H3K4me3 genes which gained K27me3+ (K4/K27+ at t=24h) maintained their averages expression level. Genes that carried both trimethyl marks at t=0 had an intermediate expression level compared to H3K4me3-marked (high expression) and H3K27me3 or non-marked genes (low expression) (**Figure 3A-D**; right panels). Approximately 20% of the double-marked genes either lost H3K4me3 or H3K27me3-enrichment in response to oxygen deprivation, while the majority of genes retained both marks. Loss of H3K27me3 did not affect gene-expression, but loss of H3K4me3 clearly reduced expression. In contrast, the H3K27me3-only pre-marked genes showed a strong shift toward gain of H3K4me3 in hypoxic conditions (±55%; **Figure 3C**). However, gain of H3K4me3 did not parallel increased gene expression as all genes remained transcriptionally silent. Of all non-marked genes (t=0) ±81% maintained this status after 24

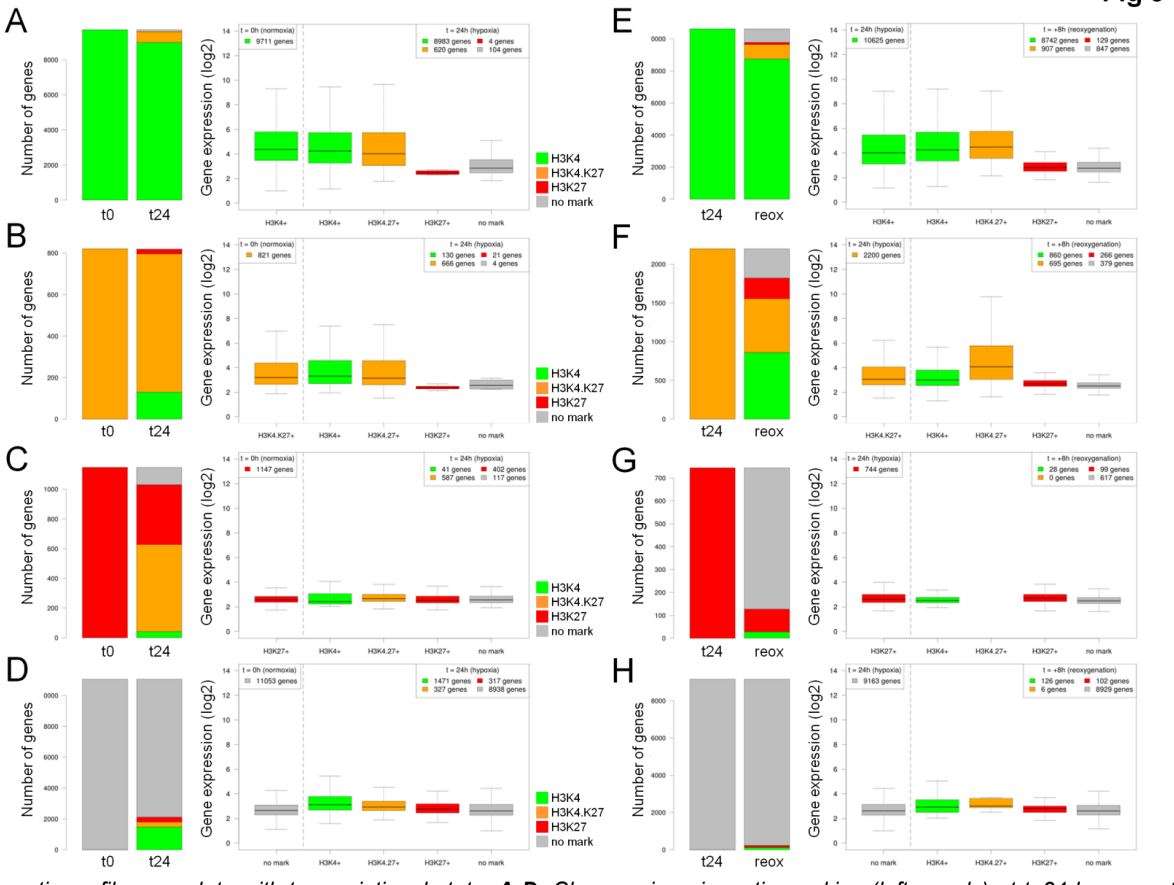


Figure 3. *Epigenetic profiles correlate with transcriptional state.* ***A-D;*** *Changes in epigenetic marking (left panels) at t=24 hours vs t=0 of genes that were* ***A****) H3K4me3-marked,* ***B****) double H3K4me3/H3K27me3–marked,* ***C****) H3K37me3-marked or* ***D****) not marked under normoxic conditions (t=0). Altered median gene expression in relation to changes in epigenetic marking (right panels).* ***E-H*** *Changes in epigenetic marking (left panels) at reoxygenation vs t=24 of genes that were* ***E****) H3K4me3-marked,* ***F****) double H3K4me3/H3K27me3–marked,* ***G****) H3K27me3-marked or* ***H****) not marked under hypoxic conditions (t=24). Altered median gene expression in relation to changes in epigenetic marking (box-plots; right panels).*

hours of hypoxia; genes that gained H3K4me3-single or both me3-marks (around 15% in total) displayed slightly elevated median expression of H3K4me3/single-marked genes; this mirrored the effect of H3K4me3-marking under oxygenated conditions, albeit with substantially lower expression levels (**Figure 3D**).

In the context of reoxygenation, the majority of H3K4me3-marked genes maintained some level of H3K4me3-marking, and gene expression was not changed upon acquisition of H3K27me3 (**Figure 3E**); non-marked genes (t=24 hrs) were slightly induced upon gain of H3K4me3 (independent of simultaneous gain of K27me3; **Figure 3H**). The double-marked and the H3K27me3-only marked genes showed the most significant loss of H3K27me3 upon reoxygenation: 56% and 87%, respectively, which paralleled the global loss of H3K27me3 (**Figure 3F,G**, *cf.* **Figure 1B**). Approximately 30% of the double-marked genes was significantly expressed at t=+8 hours (**Figure S5B**). Double-marked genes that lost H3K4me3 (30% and/or H3K27me3) showed a concomitant reduction of gene expression (**Figure 3F**). Remarkably, genes that maintained their double–marked status, showed higher transcription levels, whereas those that retained H3K4me3-only did not change their expression (**Figure 3F**), suggesting that gain of H3K27me3 is regulatory. H3K27me3–only enrichment (t=24 hrs) correlated with a transcriptionally silenced status that did not change in response to reoxygenation (**Figure 3G**).

Thus, at any given time point H3K27me3-only or no-marking generally correlated with transcriptional silencing, whether the condition was pre-existent or acquired, and gain of H3K4me3 at these sites did not meaningfully change expression. H3K4me3 pre-marked genes were generally transcriptionally active, and maintained their expression level independent of gain of H3K27me3, as long as K4 marking was not lost. Changes of H3K4me3-trimethylation primarily control transcription in the context of adaptation to altered oxygenation. In combination with H3K4me3, the H3K27me3-mark represented an important exception to the general trend that H3K27me3-marking correlates with silencing. The combined data identify H3K27me3 as the most transitory mark in response to hypoxia/reoxygenation appeared to be H3K27me3.

Normoxic trimethylation profiles correlate with transcription status

Since the (re)distribution of H3K27me3-enrichment appeared most affected by hypoxia/reoxygenation, we next determined the relation between the specific intragenic location of H3K27me3-marks and transcriptional regulation. For this purpose, the enrichment–profile of all genes was differentiated into three separate regions: promoter (promoter; -3000/-100 bp TSS), transcription start site (TSS; -100/+1000 bp), and gene-body (broad; +1000/last exon) as was defined before [37]. Genes were then assigned to one of three classes at each time point based on their H3K27me3-enrichment (**Figure S6A/a**): a distinct H3K27me3 peak upstream of the TSS (promoter class), ii) a distinct peak at the TSS (TSS–class) or iii) no peak, but instead the typical Polycomb-repression-associated ‘blanketed’ distribution over the gene body (broad class; **Figure S6Aa**). In control samples (normoxia) the majority of genes (±55%) displayed the broad (gene body/blanketing) enrichment, whereas the TSS (±22%) and promoter (±23%) class were

equally represented (**Figure 4Aa**). Hypoxia induced a 1.7 fold increase in H3K27me3-enrichment (total), and a clear shift toward TSS-class marking (±22%, t=0 to ±33%, t=8), whereas relative promoter–class marking decreased (±23%, t=0 to ±14%, t=8); this relative distribution was maintained for the duration of hypoxic exposure (**Figure 4B,C/a**). TSS-class marking decreased in response to reoxygenation (±31% to ±16%), whereas the percentage of H3K27me3-blanketed genes increased to ±75% (±55%, t=0, 24) (**Figure 4Da**). This data indicated that hypoxia induces a reversible selective increase of H3K27me3-marking at transcription start sites.

To gain insight into the biological relevance of H3K27me3-marking, average gene expression measurements were plotted for each individual subclass, and compared to the overall expression of all H3K27me3-enriched genes and non-marked genes. Distinctive TSS-marking and gene body-enrichment correlated with low expression/repression, compared to the total pool of H3K27me3-marked genes (**Figure S6A-D/b**). Promoter H3K27me3-enrichment was associated with a significantly higher average gene expression level, compared to the total H3K27me3-marked gene population, and equaled expression levels of non-marked genes; consistent with the absence of transcriptional repression, gene body H3K27me3-marking was low at promoter-marked genes (**Figure S6Aa,b)**. A similar correlation was observed between intragenic enrichment-profile and expression at acute or chronic hypoxia (**Figure S6B-C/a,b**).

We next asked whether the presence of specific H3K4me3-enrichment within these 3 H3K27me3-classes could explain the observed increased expression at these loci. To this end, all H3K27me3-marked loci were divided into two categories: HK4me3-negative and H3K4me3-marked loci (**Figure 4A-D/a**). At normoxic conditions, all three H3K27me3-classes contained bivalent genes; the highest number of double-marked genes (377; ±40%) was found in the gene body H3K27me3-class; whereas within the promoter-class loci, nearly 2/3rd of all loci was marked (253 genes) (**Figure 4Aa**). Expression within these H3K27me3-classes positively correlated with the presence of the H3K4me3-mark (**Figure 4Ac;** red boxes marked *"all"*), and with relative H3K4me3/TSS-enrichment (promoter and gene-body class; **Figure 4Ab**); TSS-double marked genes showed the lowest median expression. Under hypoxic conditions, the number of double-marked genes increased substantially in each H3K27me3-class (**Figure 4B,C/a**), yet the normoxic expression-ratio was maintained (*i.e.* promoter class > gene body class > TSS class), despite progressive reduction of H3K4me3-marking at the TSS in all three H3K27me3-classes (**Figure 4B-C/b,c**).

The association between intragenic H3K27me3-marking and transcriptional regulation was lost in the context of reoxygenation stress (**Figure S6Da,b**; *cf.* **Figure 2Bc**); of note: even silenced loci (TSS-peak or blanketing-type distribution; repressed at t=0 and during hypoxia), showed a sudden rise of transcriptional activity (**Figure S6Db**). Loss of repression under these conditions was not explained by quantitative changes in H3K27me3-enrichment (data not shown). Following reoxygenation, expression was induced at all H3K4me3-marked loci irrespective of H3K27me3-subclass, but not at H3K27me3/single-marked loci (**Figure 4Da-c**). Of note, the relatively highest increase of H3K4me3-

Fig 4

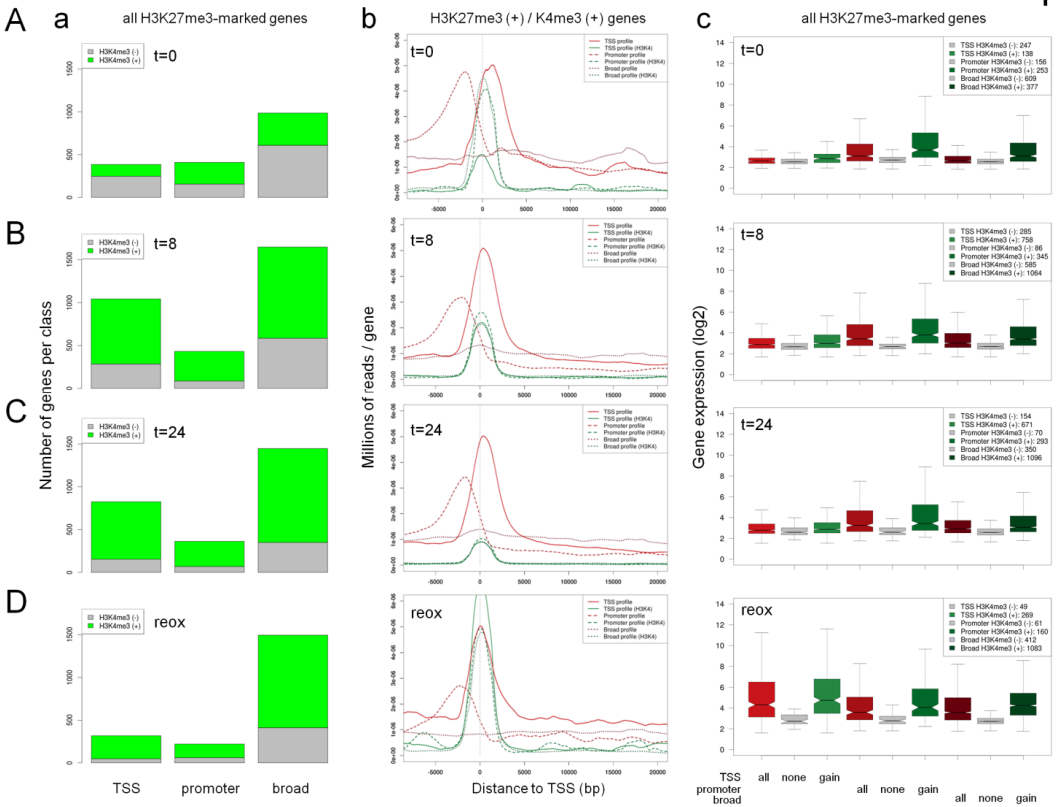


Figure 4. *Transcription status is primarily controlled by H3K4me3.* ***A-D****;* ***a*** *panels) H3K27me3-enriched genes were classified as TSS-, promoter- and gene body (broad)-marked genes and differentiated according to their H3K4me3-status (marked (+) or non-marked (-);* ***b*** *panels) TSS-associated enrichment profiles (average number of reads) for the corresponding H3K27me3-classes (i.e. +/- gain of H3K4me3 as in a panel); red lines (solid/dashed/dotted) represent all single H3K27me3-marked genes, green lines represent all double H3K4me3/K27me3-marked genes (sold/dashed/dotted lines correspond to H3K27me3 classes in (a);* ***c*** *panels) box-plots: gain of H3K4me3-enrichment was plotted against median expression levels. Comparative analysis was done at (A) t=0, (B) t=8, (C) t=24 hours hypoxia and (D) after reoxygenation (t=+8 hours). Bottom legend (c): subscripts correspond to all H3K27me3-marked genes (all), the H3K27me3-only (none) and H3K4me3/H3K27me3-marked sequences (gain), and are differentiated between promoter, TSS and gene body classes as indicated.*

occupation within the H3K27me3/TSS-class correlated with the highest median expression induction (3,7 fold change; t=24 *vs.* reox).

Thus, our comparative analysis of intragenic marking and gene activity identified H3K4me3-marking as the dominant determinant of expression of a locus. The data suggested that transcriptional status under hypoxic conditions was refractory to gain of H3K27me3-enrichment and revealed a clear association between defined promoter-type H3K27me3-marking and a transcriptionally permissive state (*vs.* TSS or gene body-enrichment and transcriptional repression) under normoxic and hypoxic conditions. Reoxygenation induced massive transcriptional deregulation, irrespective of the H3K27me3-status of loci.

Hypoxia-induced bivalency partially overlaps with bivalent genes in ES-cells

More than 800 genes were classified as bivalently marked under normoxic conditions, based on co-occurrence of H3K4me3 and H3K27me3-marks within genic regions (**Figure 3B,F**). To determine whether bivalent genes specifically mapped to any of the intragenic H3K27me3-profile subclasses, we analyzed the promoter, TSS or gene body H3K27me3-marking groups for H3K4me3-presence. Approximately 80% (345 genes) of H3K27me3/promoter-enriched genes was also enriched for H3K4me3 in hypoxic samples (*cf.* **Figure 4A-D/a**). The broadly H3K27me3-enriched genes showed a relatively lower H3K4me3-enrichment (±64%,1064 genes; t=8 hrs) compared to promoter or TSS-class genes (758 genes); of TSS-H3K27me3 genes ±72% was associated with H3K4me3, (*cf.* **Figure 2Ab**); the presence of H3K4me3 around the TSS (**fig** *cf.* **fig 2**) determined gene expression status, independent of the intragenic location of K3K27me3. Consistent with the earlier observed transcriptional deregulation at reoxygenation, like H3K27me3, H3K4me3-occupation failed to correlate with transcriptional status. This data indicated that bivalent genes are strongly represented among H3K27me3/TSS and promoter-marked genes.

To determine how bivalency was accomplished, the original epigenetic status of the bivalently marked genes (at t=8 hrs hypoxia) was traced back to normoxia and displayed as a function of time spent at low oxygen. We first selected the TSS-associated bivalency (H3K4me3 and H3K27me3 at TSS) genes for further analysis. Bivalency increased approximately 3 fold (758 to 2157 genes) in response to loss of oxygen (**Figure 5Aa**). Gain of H3K4me3 appeared under-represented in regards to acquisition of bivalency, due to the fact that a many TSS genes had pre-existent H3K4me3-marking at baseline. More than half of the bivalent genes (±54%; 1173 genes) at 8 hours of hypoxia had acquired H3K27me3-marking alone or in combination with H3K4me3-marking in response to oxygen withdrawal (**Figure 5Aa**). During prolonged hypoxia (t=8 to t=24 hours) or at reoxygenation, again more than ±50% of acquired bivalent loci showed gain of H3K27me3 (**Figure 5Ba**); this finding is in good agreement with the above proposed notion that H3K27me3-marking is most prone to dynamic modulation.

Expression levels of genes with acquired bivalency, largely reflected those corresponding to their original epigenetic status: for instance, genes that gained H3K27me3 were significantly higher expressed compared to stable bivalent genes (**Figure 5Ac,Bc**). Although a possible contribution of mRNA stability

Fig 5

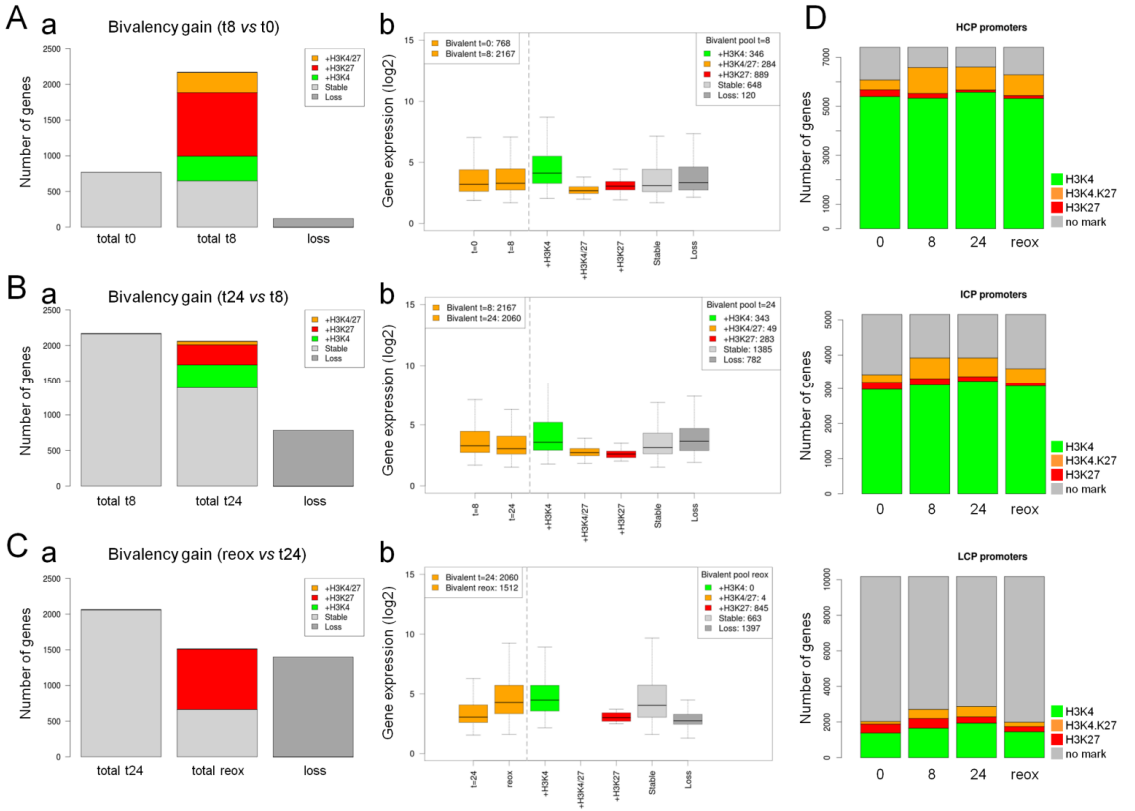
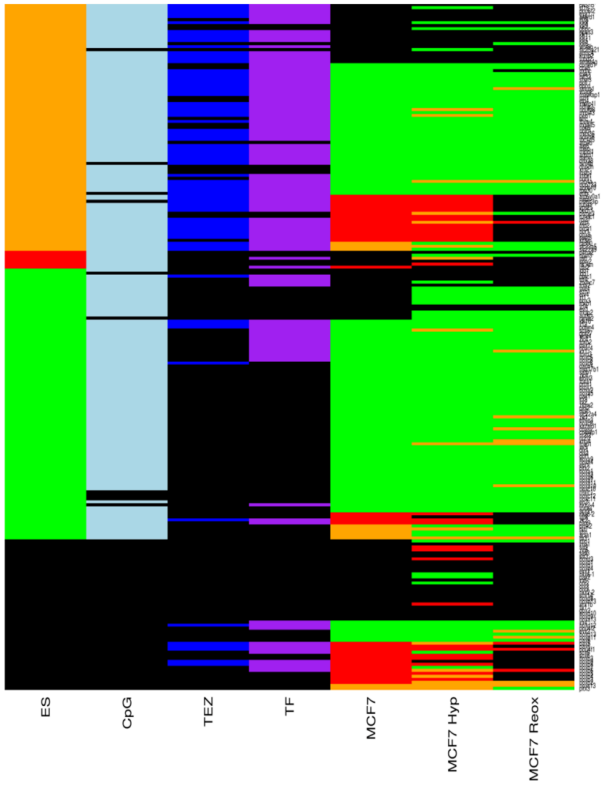


Figure 5. *Hypoxia-induced bivalency.* ***A-C****;* ***a*** *panels) Gain of epigenomic bivalency between* ***A****) control cells and hypoxic cells (t=8 hrs);* ***B****) during hypoxia (t=24 vs t=8 hrs hypoxia) and* ***C****) upon reoxygenation (reox vs t=24 hrs hypoxia); indicated are gene categories that acquired bivalency by: gain of H3K4me3, gain of H3K27me3, gain of both, or that had retained bivalency; left bars indicate bivalency at previous time point; bars to the right indicate loss of bivalency at time point analyzed;* ***b*** *panels) box-plots indicating median expression values of bivalently marked genes at indicated experimental time points; indicated are gene categories that acquired bivalency by: gain of H3K4me3, gain of H3K27me3, gain of both, or that had retained bivalency; indicated are also sites at which bivalency was lost (far right boxes);* ***D****) distribution of H3K4me3 and H3K27me3-marks at (top) high-CpG, (middle) intermediate and (bottom) low CpG-content promoters.*

could formally not be excluded, transcriptional changes should be detectable over an 8 hrs interval and, moreover, we observed a similar trend for genes that gained H3K4me3.
Upon reoxygenation many bivalent genes were lost, and despite a substantial gain of bivalency, the number of bivalently marked loci dropped to ±75% compared at t=24 and was still 2 fold higher compared to normoxia: more than 500 loci displayed bivalent epigenetic marking at the reoxygenation time point, suggesting that somehow these genes were shielded from acute HKDM activity (**Figure 5Ca**). Compared to loss of H3K4me3-marking in response to reoxygenation (±9% of H3K4me3/single-marked genes; t=24 hrs *vs*. reox), H3K27me3 was considerably less stable (±87% at H3K27me3/single-marked genes). At bivalent loci, the loss was relatively larger for H3K4me3 (±29%), whereas H3K27me3-loss was substantially less (± 56%) compared to single-marked genes.
In conclusion, hypoxia induces epigenetic bivalency, which correlates best with a TSS-specific increase of H3K27me3. Bivalent genes show transcriptional activity which overall resembles their original expressional status under normoxia.

Bivalency in embryonal stem cells (ESC) coincides with genomic CpG-content (*i.e.* CpG islands) [42]. To verify whether this applied to bivalent epigenetic marking in MCF7 cancer cells as well, the prevalence of bivalently marked genes in high CpG-content promoter/enhancer regions (HCP) *versus* intermediate and low CpG-regions (ICP and LCP, respectively) was determined. H3K4me3 preferentially localized to HCP regions; 5198 HCP, 2775 ICP and 1015 LCP promoters were positive for H3K4me3 at normoxia, corresponding to 70% of charted HCPs, 54% of ICPs and 10% of LCPs respectively (**Figure 5D**). Bivalently-marked genes showed a highly similar distribution pattern: 366 HCP, 183 ICP and 72 LCP promoters were positive for both trimethyl marks in control cells (t=0). Relatively few promoters showed H3K27me3-marking, consistent with the relative H3K27me3-enrichment over gene bodies or at the TSS. H3K27me3/promoter distribution was opposite of that of H3K4me3: 310 HCP, 2720 ICP and 577 LCP promoters were enriched for H3K27m3; (**Figure 5D,** *cf.* **Figure 4**). Although the overall number of H3K4me3, H3K27me3 or H3K4me3/H3K27me3-enriched genes increased in response to hypoxia, their relative association to HCP, ICP and LCP was maintained (data not shown).
In the context of stem/progenitor cells bivalently marked loci are often found among key developmental control genes [11-13]. We next examined whether the hypoxia-induced bivalently marked gene set in MCF7 cancer cells showed overlap with bivalent genes previously identified in ESC. Approximately 600 of the hypoxia-induced bivalent genes in MCF7 cells match a previously identified group of bivalent genes in ESC [12] (**Figure 6A**). Comparative GO analysis of both bivalent gene pools revealed a distinctive overlap of functional pathways involved and confirmed the presence of bivalently marked developmental genes (**Figure 6B**). Many of these genes were already bivalently marked at the 8 hour hypoxia time point (**Figure S7**), suggesting that these loci are pre-set targets for bivalent marking under low oxygen. Hence, our findings suggest that oxygen deprivation in breast cancer cells induces bivalent epigenetic marking at genes which are also bivalently marked in ESC and which control key processes during development.

A



B

Bivalency 24 hours hypoxia

GO term	# genes
multicellular organismal development	390
signal transduction	374
regulation gene expression	224
nervous system development	197
intracellular signal transduction	172
programmed cell death	141
neuron differentiation	124
protein phosphorylation	110
response to wounding	106
cell adhesion	97
cell migration	82
central nervous system development	72
blood coagulation	64
pos regulation apoptosis	59
neg regulation apoptosis	57
pos regulation transcription by RNApol II	53
axon guidance	51
brain development	49
neg regulation transcription by RNApol II	47
actin cytoskeleton organization	42
morphogenesis epithelium	41
regulation cell migration	41
response to drug	36
angiogenesis	34
platelet activation	34
response to metal ion	30
pos regulation cell migration	30
cellular amino acid metabolic process	28
kidney development	27
NGFR signaling pathway	27
osteoblast differentiation	25
morphogenesis a branching structure	25
leukocyte migration	22
telencephalon development	21
regulation of MAP kinase activity	21
cellular response to growth factor stimulus	21
intracellular receptor signaling pathway	20
learning or memory	19
bone mineralization	19
heart morphogenesis	18
Odontogenesis	18
BMP signaling pathway	17
bone development	16
activation phospholipase C activity	15
pos regulation MAPKKK cascade	15
embryonic skeletal system development	15
ureteric bud development	14
embryonic limb morphogenesis	14
regulation bone mineralization	14
activation MAPK activity	13
regulation peptidyl-tyrosine phosphorylation	13
Neg regulation Wnt-R signaling	13
chromatin remodeling	12
pos regulation epithelial cell proliferation	12
chondrocyte differentiation	11
regulation transcription RNApol II by nuHR	10

Bivalency 24 hours hypoxia (cont'd)

GO term	# genes
Integrin-mediated signaling pathway	9
positive regulation bone mineralization	9
odontogenesis dentine-containing tooth	9
pos regulation osteoblast differentiation	9
VEGFR signaling pathway	9
regulation dendrite development	9
neg regulation T cell activation	9
response to interleukin-1	9
cellular response to retinoic acid	9
mesoderm formation	8
female gonad development	8

Bivalency ESC

GO term	# genes
multicellular organismal development	93
regulation transcription, DNA-dependent	66
organ morphogenesis	49
transcription by RNApol II	46
pattern specification process	46
embryo development	44
pos regulation transcription, DNA-dpt	38
neuron differentiation	34
skeletal system development	33
central nervous system development	30
embryo development	30
pos regulation RNA pol II transcription	30
anterior/posterior pattern specification	27
neg regulation transcription, DNA-dpt	25
brain development	21
embryonic organ morphogenesis	21
neg regulation RNA pol II transcription	20
skeletal system morphogenesis	20
embryonic skeletal system development	20
cell proliferation	19
neuron development	17
embryonic skeletal system morphogenesis	15
blood vessel morphogenesis	12
dorsal/ventral pattern formation	11
neuron fate specification	11
camera-type eye development	10
regulation DNA binding TF factor activity	10
kidney development	9
limb morphogenesis	9
cartilage development	9
angiogenesis	8
forebrain development	8

Figure 6. *Comparative analysis of hypoxia-induced bivalent genes in MCF7 cells and bivalent loci in ESC.* ***A****; Heatmap indicating overlap of bivalent genes in this study with bivalently marked genes in embryonal stem cells (ESC; orange = bivalently marked, red = H3K27me3 only, green = H3K4me3 only, black = no marking), transcription factor genes (TF; blue = TF gene, black = not a TF gene), genes containing promoter CpG islands (CpG; blue = CpG island present, black = no CpG island) and genes whose TSS coincides with transposon-exclusion zones (TEZ; blue = located in TEZ, black = not TEZ associated).* ***B****; Gene Ontology enrichment analysis of bivalent genes; listed are GO analysis at t=0, t=8 and t=24 hours hypoxia; same GO-analysis was done on ESC-bivalent markers published elsewhere (Bernstein et al., 2006).*

Discussion

We here report that oxygen deprivation induces massive genome-wide trimethylation at H3K4 and H3K27 through inhibition of JHDM function. Our data establish that oxygen-sensing by HKDM represents a direct link between the micro-environment and epigenetic regulatory mechanisms. Comparative analysis of ChIP-deep sequencing and expression array data sets revealed that H3K27me3-occupation at the TSS and over the gene body correlates with transcriptional repression, whereas H3K4me3 marking under all conditions correlates with gene activity. We establish H3K4me3-enrichment as the most relevant determinant of transcription status. H3K27me3-acquisition at the TSS with pre-existent H3K4me3 marking does not alter the expression status. We show that a subset of genes acquired bivalency, by gain of H3K27me3 (54% 1173 genes). The bivalent subset represents genes previously identified as bivalently marked genes in ESC.

Recent advances in the field of epigenetics suggest that chromatin state is more dynamic than originally anticipated. Comprehensive insight into transcriptional reprogramming in response to changing microenvironments requires systematical genome-wide mapping of epigenetic marks as a function of cell type, differential state and micro-environment [6]. We here studied dynamic changes in histone trimethylation in the context of cellular adaptation to changed oxygenation (hypoxic stress, reoxygenation). To robustly detect H3K27me3-enrichment, we designed a novel enrichment-finding protocol including a normalization/summarization strategy. Importantly, this also allows us to obtain quantitative measurements of H3K27me3-marking between samples (time-points). Applying this novel strategy we were able to reproduce reported correlation between gene-body associated 'blanketing' H3K27me3-profile and transcriptional repression, thereby validating our analytical approach [33]. The prominent TSS-centered H3K4me3-enrichment at expressed loci, and the marked low nucleosome abundance right over the TSS, is also consistent with earlier findings [11].

We detected a clear positive correlation between TSS/H3K4me3-enrichment and gene expression at all time-points studied. H3K27me3 displayed multiple distinct enrichment profiles, in line with previous publications [37, 43]. H3K27me3-enrichment across the gene body followed the classical Polycomb-associated repressive H3K27me3-enrichment profile, which is referred to as "blanketing" [31, 32]. Specific enrichment at the promoter region, in combination with a marked depletion of the signal within the gene body, overlaps with H3K4me3-enrichment profiles and appears permissive for active transcription; consequently bivalently-marked genes are expressed at low levels [12, 13, 31]. Interestingly, we observed enhanced H3K27me3-marking upon oxygen deprivation at the TSS, relative to promoter and gene-body; this resulted in bivalent marking of a substantial number of genes (±950-1100; t=8 and 24 hrs combined). By inference, inhibition of HKDMs is likely to be responsible for the increased bivalency at TSS in the absence of molecular oxygen. The co-occurrence of both methyl marks was first described in embryonic stem cells and is thought to mark key developmental control genes as "poised" for transcriptional activation. We and others found a marked increase of bivalent marking at CpG-rich genomic regions [37]. CpG-islands are usually hypomethylated (at the DNA level), but instead are marked

by trimethylation on histone 3 lysine 4 (H3K4me3) or lysine 27 (H3K27me3) [10, 37]; conversely, bivalent domains highly correlate with the presence of CpG-islands, and may reflect a competition between PcG-recruitment (silencing) and transcriptional activation [44]. A recent study in *Xenopus* showed that bivalency may, in reality, affect different alleles [45]. Hence, to conclusively prove that H3K4me3 and H3K27me3 occur on the same chromatin segments and are not merely a reflection of different cell pools, sequential ChIP for both marks needs to be performed.

Enhancers represent an important class of functional domains involved in transcriptional regulation. In contrast to promoter regions, enhancers operate as highly tissue-specific elements, and their (functional) existence cannot be directly inferred from underlying DNA-sequences [46]. In addition, enhancers may be located thousands of base pairs removed *in cis* from their target genes, or even on a different chromosome, which makes it difficult to define and relate such regulatory regions to genes. No public MCF7 ChIP-seq data-set is currently available, which includes TF/DNA-binding profiles and/or enhancer-associated epigenetic regulatory factors (e.g. p300) or associated histone modifications (e.g. H3K4me1 and H3K27ac). To probe for such histone modifications would be relevant in light of our current findings, as it was reported that bivalent promoters are associated to a specific subclass of enhancer elements which are H3K4me1/H3K27me3-marked instead of H3K4me1/H3K27ac which is the case for active promoters [47].

The increased global methylation we observed could either be the result of increased histone methyl transferase activity or reduced histone demethylase activity. Transcriptional analysis of candidate enzymes involved in histone (de)methylation did not suggest a singular role for any of these factors in this respect. Expression of JHDM proteins is known to be controlled by HIF1a; likewise HTM expression is controlled by oxygen. Nevertheless, we showed that the hypoxia-increased trimethylation was HIF1a–independent. Most demethylases depend on molecular oxygen for their activity, making JHDM oxygen sensors

The exact relevance of Polycomb-associated marking and occupation at different intragenic positions is currently not fully understood. The H3K27me3-mark is bound by chromobox proteins belonging to Polycomb group proteins [48]. PcG proteins were shown to be present at active RPOL2 promoters in *Drosophila* and to interact with basal transcription factors (TF) [49]. Target-gene silencing via distant enhancer-looping to gene promoters has been shown in flies and in mammalian systems [50]. In addition, PcG complexes associate with splicing factors, revealing an as of yet poorly understood role in regulation of gene expression. In the context of our hypoxic-stress model, increased bivalency mostly correlated with gain of TSS-H3K27me3. It is also interesting to note that H3K27me3-marking appeared to be the most prone to dynamic change. The transcriptional status of the corresponding loci, however, was generally maintained; hence, this finding suggests that, once marked for transcription, recruitment of HMT activity and concomitant H3K27me3-marking *per se* is not sufficient for transcriptional repression.

Despite the global demethylation following reoxygenation, a substantial number of loci was bivalently marked 8 hours into restoration of oxygen levels. It is currently unclear whether these loci are protected from HKDM activity and/or selectively methylated by K27-directed HTMs. It is conceivable that some other local epigenetic aspect (e.g. inhibitor recruitment of higher-order chromatin structure) prevents access and/or shields nucleosomes against demethylation. Of relevance, HKMTs are often found in close conjunction with HKDMs for opposing epigenetic marks; together these paired epigenetic modulators are thought to reinforce transcriptional decisions (co-stability) [51]. It is tempting to speculate about a causal role for similarly cooperative H3K4me3 HMTs/H3K27me3 HDMs in the establishment of bivalency under HKDM-inhibitory conditions. Irrespective of the exact underling mechanism, our findings confirmed our initial hypothesis and define a novel role for oxygen in epigenetic regulation of stress responses. Future experiments should address the question whether loss of UTX/JMJD3 abolishes the increase in H3K27me3 levels, both on the global as well as on individual gene level. Although the involvement of demethylases was recently addressed in the context of hypoxia-induced increased H3K4me3-marking [40], our studies provide additional insight in to the effects of reoxygenation on epigenomic and transcriptomic responses.

Acquisition of bivalency under hypoxic conditions is a phenomenon which may have relevance in the context of tumor biology (*i.e.* plasticity, malignancy). TSS-associated H3K27me3 was previously shown to be prominent in embryonic stem cells [37]. As bivalency was proposed to represent an aspect of hierarchical differentiation, increased epigenomic bivalency may reflect acquisition of a more primitive chromatin state [8, 52, 53]. Of relevance, stem cell niches are known to be hypoxic [54-57]. As cancer development is thought to be sustained by cancer stem cells [58-61], it is tempting to speculate about a possible role for bivalent marking in re-establishing a 'poised' gene status, and that hypoxic microenvironment selects for or drives tumor plasticity through acquisition of a less differentiated epigenome. In this study, we applied severe hypoxic conditions (<0.02% O_2) to achieve a semi-synchronized cell population response. As tumor cells are subjected to constantly fluctuating oxygen concentrations due to poor vasculature, the effect of exposure to varying hypoxic as well as cycling hypoxia-reoxygenation conditions on epigenomic remodeling and associated transcriptional changes needs to be addressed [62]. Repetitive exposure to hypoxia/reoxygenation-stress may impose a more realistic micro-environment and concomitant selection-pressure on a cancer cell population and support increased tumor plasticity and malignant progression. Our observations put forward the exciting possibility that repeated exposure to hypoxia may promote or select for stem-ness or cancer stem cell phenotypes.

To our knowledge, our study for the first time combines analysis of dynamic data sets on gene expression and histone methylation in the context of adaptation to acute environmental changes. The number of studies on altered chromatin states and the role of epigenetic modifiers, *e.g.* in the context of embryogenesis, spermatogenesis, metabolic disorders and tumorigenesis steadily increases [52, 63]. Combined epigenomic profiling and detailed analysis of transcriptional reprogramming will be crucial for fundamental understanding of pluripotency and clinical application of induced precursor cells.

Acknowledgements

We are much indebted to many colleagues (*cf.* Methods section) for sharing research materials, to Caroline Gits for technical support of the study, to Joep Geraedts and Ronit Sverdlov for critically reading the manuscript and to Chris Evelo, Lars Eijssen, Timothy Beck, Marianne Koritzinsky, members of the MAASTRO and Molecular Genetics departments for scientific discussions. These studies received financial support from the Dutch Science Organization (ZonMW-NWO): VIDI grant 016.046.362 (JWV); transnational University Limburg (tUL) grant (JWV/BGW).

References

1. Brown JM: The hypoxic cell: a target for selective cancer therapy--eighteenth Bruce F. Cain Memorial Award lecture. Cancer Res 1999, 59(23):5863-5870.
2. Kenneth NS, Rocha S: Regulation of gene expression by hypoxia. Biochem J 2008, 414(1):19-29.
3. Semenza GL: Targeting HIF-1 for cancer therapy. Nat Rev Cancer 2003, 3(10):721-732.
4. Brizel DM, Sibley GS, Prosnitz LR, Scher RL, Dewhirst MW: Tumor hypoxia adversely affects the prognosis of carcinoma of the head and neck. Int J Radiat Oncol Biol Phys 1997, 38(2):285-289.
5. Hockel M, Schlenger K, Aral B, Mitze M, Schaffer U, Vaupel P: Association between tumor hypoxia and malignant progression in advanced cancer of the uterine cervix. Cancer Res 1996, 56(19):4509-4515.
6. Berger SL: The complex language of chromatin regulation during transcription. Nature 2007, 447(7143):407-412.
7. Kouzarides T: Chromatin modifications and their function. Cell 2007, 128(4):693-705.
8. Bernstein BE, Meissner A, Lander ES: The mammalian epigenome. Cell 2007, 128(4):669-681.
9. Esteller M: Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet 2007, 8(4):286-298.
10. Azuara V, Perry P, Sauer S, Spivakov M, Jorgensen HF, John RM, Gouti M, Casanova M, Warnes G, Merkenschlager M et al: Chromatin signatures of pluripotent cell lines. Nat Cell Biol 2006, 8(5):532-538.
11. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: High-resolution profiling of histone methylations in the human genome. Cell 2007, 129(4):823-837.
12. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K et al: A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 2006, 125(2):315-326.
13. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP et al: Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 2007, 448(7153):553-560.
14. Cao R, Wang L, Wang H, Xia L, Erdjument-Bromage H, Tempst P, Jones RS, Zhang Y: Role of histone H3 lysine 27 methylation in Polycomb-group silencing. Science 2002, 298(5595):1039-1043.

15. Swigut T, Wysocka J: H3K27 demethylases, at long last. Cell 2007, 131(1):29-32.
16. Schuettengruber B, Chourrout D, Vervoort M, Leblanc B, Cavalli G: Genome regulation by polycomb and trithorax proteins. Cell 2007, 128(4):735-745.
17. Brock HW, Fisher CL: Maintenance of gene expression patterns. Dev Dyn 2005, 232(3):633-655.
18. Cloos PA, Christensen J, Agger K, Helin K: Erasing the methyl mark: histone demethylases at the center of cellular differentiation and disease. Genes Dev 2008, 22(9):1115-1140.
19. Hansen KH, Bracken AP, Pasini D, Dietrich N, Gehani SS, Monrad A, Rappsilber J, Lerdrup M, Helin K: A model for transmission of the H3K27me3 epigenetic mark. Nat Cell Biol 2007, advanced online publication.
20. Chan DA, Giaccia AJ: Hypoxia, gene expression, and metastasis. Cancer Metastasis Rev 2007, 26(2):333-339.
21. Johnson AB, Denko N, Barton MC: Hypoxia induces a novel signature of chromatin modifications and global repression of transcription. Mutat Res 2008, 640(1-2):174-179.
22. Hou H, Yu H: Structural insights into histone lysine demethylation. Curr Opin Struct Biol 2010, 20(6):739-748.
23. Lohse B, Kristensen JL, Kristensen LH, Agger K, Helin K, Gajhede M, Clausen RP: Inhibitors of histone demethylases. Bioorg Med Chem 2011, 19(12):3625-3636.
24. Selak MA, Armour SM, MacKenzie ED, Boulahbel H, Watson DG, Mansfield KD, Pan Y, Simon MC, Thompson CB, Gottlieb E: Succinate links TCA cycle dysfunction to oncogenesis by inhibiting HIF-alpha prolyl hydroxylase. Cancer Cell 2005, 7(1):77-85.
25. Shi Y, Whetstine JR: Dynamic regulation of histone lysine methylation by demethylases. Mol Cell 2007, 25(1):1-14.
26. Smith ER, Lee MG, Winter B, Droz NM, Eissenberg JC, Shiekhattar R, Shilatifard A: Drosophila UTX is a histone H3 Lys27 demethylase that colocalizes with the elongating form of RNA polymerase II. Mol Cell Biol 2008, 28(3):1041-1046.
27. Ivan M, Kondo K, Yang H, Kim W, Valiando J, Ohh M, Salic A, Asara JM, Lane WS, Kaelin WG, Jr.: HIFalpha targeted for VHL-mediated destruction by proline hydroxylation: implications for O2 sensing. Science 2001, 292(5516):464-468.
28. Jaakkola P, Mole DR, Tian YM, Wilson MI, Gielbert J, Gaskell SJ, Kriegsheim A, Hebestreit HF, Mukherji M, Schofield CJ et al: Targeting of HIF-alpha to the von Hippel-Lindau ubiquitylation complex by O2-regulated prolyl hydroxylation. Science 2001, 292(5516):468-472.
29. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A et al: Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 2007, 4(8):651-657.
30. Park PJ: ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 2009, 10(10):669-680.

31. Bracken AP, Dietrich N, Pasini D, Hansen KH, Helin K: Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. Genes Dev 2006, 20(9):1123-1136.
32. Pauler FM, Sloane MA, Huang R, Regha K, Koerner MV, Tamir I, Sommer A, Aszodi A, Jenuwein T, Barlow DP: H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. Genome Res 2009, 19(2):221-233.
33. Adriaens M, Prickaerts P, Chan-Seng-Yue M, Beck T, Wouters BG, Voncken JW, Evelo C: Capturing ChIP-seq profiles of H3K27me3 in dynamic biological systems. submitted.
34. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Forrest Spencer F: A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. J Am Stat Assoc 2004, 99(247):909-917.
35. Alexa A, Rahnenfuhrer J, Lengauer T: Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 2006, 22(13):1600-1607.
36. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000, 25(1):25-29.
37. Young MD, Willson TA, Wakefield MJ, Trounson E, Hilton DJ, Blewitt ME, Oshlack A, Majewski IJ: ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. Nucleic Acids Res 2011, 39(17):7415-7427.
38. Krieg AJ, Rankin EB, Chan D, Razorenova O, Fernandez S, Giaccia AJ: Regulation of the Histone Demethylase JMJD1A by Hypoxia-Inducible Factor 1a Enhances Hypoxic Gene Expression and Tumor Growth. Molecular and Cellular Biology 2010, 30(1):344-353.
39. Yang J, Jubb AM, Pike L, Buffa FM, Turley H, Baban D, Leek R, Gatter KC, Ragoussis J, Harris AL: The histone demethylase JMJD2B is regulated by estrogen receptor alpha and hypoxia, and is a key mediator of estrogen induced growth. Cancer Res 2010, 70(16):6456-6466.
40. Zhou X, Sun H, Chen H, Zavadil J, Kluz T, Arita A, Costa M: Hypoxia induces trimethylated H3 lysine 4 by inhibition of JARID1A demethylase. Cancer Res 2010, 70(10):4214-4221.
41. Jiang C, Pugh BF: Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet 2009, 10(3):161-172.
42. Ku M, Koche RP, Rheinbay E, Mendenhall EM, Endoh M, Mikkelsen TS, Presser A, Nusbaum C, Xie X, Chi AS et al: Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. PLoS Genet 2008, 4(10):e1000242.
43. Hon GC, Hawkins RD, Ren B: Predictive chromatin signatures in the mammalian genome. Hum Mol Genet 2009, 18(R2):R195-201.
44. Lynch MD, Smith AJ, De Gobbi M, Flenley M, Hughes JR, Vernimmen D, Ayyub H, Sharpe JA, Sloane-Stanley JA, Sutherland L et al: An interspecies analysis reveals a key role for unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment. EMBO J 2011, 31(2):317-329.

45. Akkers RC, van Heeringen SJ, Jacobi UG, Janssen-Megens EM, Francoijs KJ, Stunnenberg HG, Veenstra GJ: A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in Xenopus embryos. Dev Cell 2009, 17(3):425-434.
46. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA et al: Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 2007, 39(3):311-318.
47. Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J: A unique chromatin signature uncovers early developmental enhancers in humans. Nature 2011, 470(7333):279-283.
48. Min J, Zhang Y, Xu RM: Structural basis for specific binding of Polycomb chromodomain to histone H3 methylated at Lys 27. Genes Dev 2003, 17(15):1823-1828.
49. Breiling A, Turner BM, Bianchi ME, Orlando V: General transcription factors bind promoters repressed by Polycomb group proteins. Nature 2001, 412(6847):651-655.
50. Simon JA, Kingston RE: Mechanisms of polycomb gene silencing: knowns and unknowns. Nat Rev Mol Cell Biol 2009, 10(10):697-708.
51. Dodd IB, Micheelsen MA, Sneppen K, Thon G: Theoretical analysis of epigenetic cell memory by nucleosome modification. Cell 2007, 129(4):813-822.
52. Fisher CL, Fisher AG: Chromatin states in pluripotent, differentiated, and reprogrammed cells. Curr Opin Genet Dev 2011, 21(2):140-146.
53. Herz H-M, Nakanishi S, Shilatifard A: The Curious Case of Bivalent Marks. Developmental Cell 2009, 17(3):301-303.
54. Covello KL, Kehler J, Yu H, Gordan JD, Arsham AM, Hu CJ, Labosky PA, Simon MC, Keith B: HIF-2alpha regulates Oct-4: effects of hypoxia on stem cell function, embryonic development, and tumor growth. Genes Dev 2006, 20(5):557-570.
55. Mohyeldin A, Garzon-Muvdi Ts, Quinones-Hinojosa A: Oxygen in Stem Cell Biology: A Critical Component of the Stem Cell Niche. Cell Stem Cell 2010, 7(2):150-161.
56. Silvan U, Diez-Torre A, Arluzea J, Andrade R, Silia M, Arechaga J: Hypoxia and pluripotency in embryonic and embryonal carcinoma stem cell biology. Differentiation 2009, 78(2â€“3):159-168.
57. Welford SM, Giaccia AJ: Hypoxia and Senescence: The Impact of Oxygenation on Tumor Suppression. Molecular Cancer Research 2011, 9(5):538-544.
58. Clarke MF, Fuller M: Stem Cells and Cancer: Two Faces of Eve. Cell 2006, 124:1111-1115.
59. Hanahan D, Weinberg RA: Hallmarks of cancer: the next generation. Cell 2011, 144(5):646-674.
60. Lobo NA, Shimono Y, Qian D, Clarke MF: The Biology of Cancer Stem Cells. Annual Review of Cell and Developmental Biology 2007, 23(1):675-699.
61. Woodward WA, Chen MS, Behbod F, Rosen JM: On mammary stem cells. J Cell Sci 2005, 118(Pt 16):3585-3594.
62. Chan N, Koch CJ, Bristow RG: Tumor hypoxia as a modifier of DNA strand break and cross-link repair. Curr Mol Med 2009, 9(4):401-410.

63. Bogdanovic O, van Heeringen SJ, Veenstra GJ: The epigenome in early vertebrate development. Genesis 2011.

Supplemental Figures

Supplemental figures S1 to S7 can be found on the next pages. Supplemental methods follows thereafter.

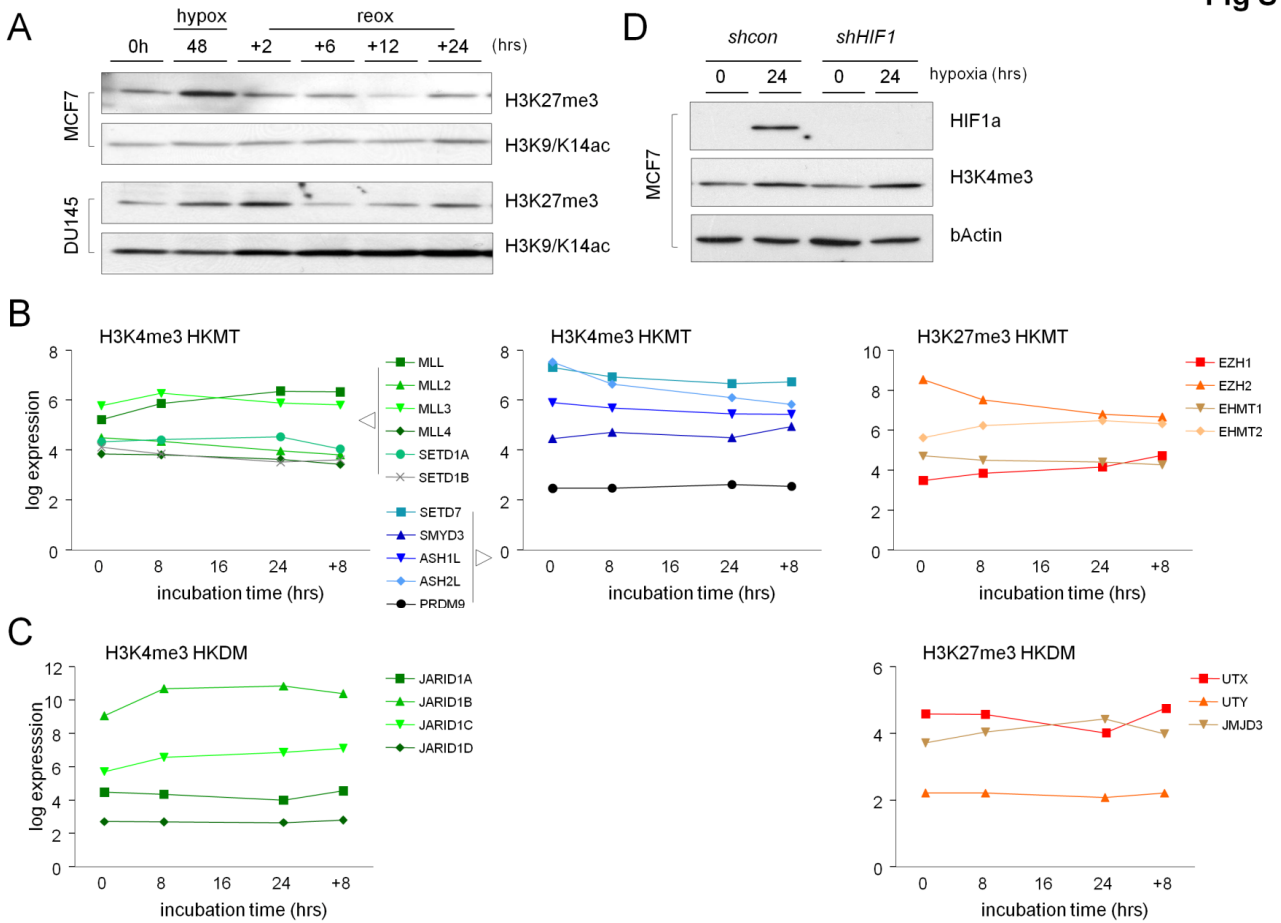


Figure S1. *Reversible oxygen-dependent global changes of H3K4me3 and H3K27me3 levels.* ***A****; Immuno-blot detection (IB) of epigenetic changes (H3K4me3 and H3K27me3) in MCF7 and DU145 cells.* ***B-C****; expression levels changes of* ***B****) known and putative H3K4 (left) and H3K27 (right) methyl transferases (HKMT);* ***C****) known and putative H3K4 (left) and H3K27 (right) demethylases (HKDM).*

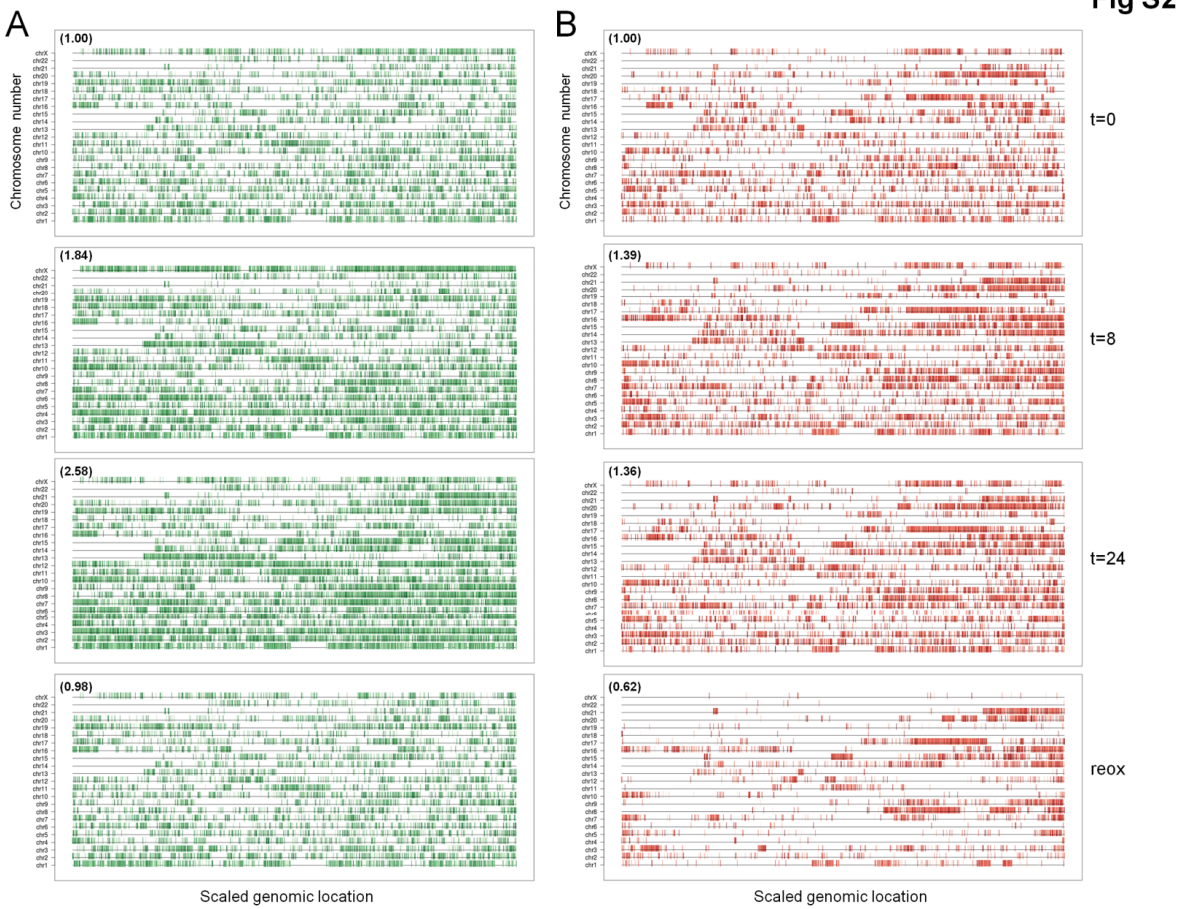


Figure S2. *Reversible oxygen-dependent global changes of H3K4me3 and H3K27me3 levels.* ***A-B****; Genome-wide presentation of* ***A****) H3K4me3 and* ***B****) H3K27me3 enrichment/peak locations at indicated experimental time points. Numbers in brackets indicate fold change in peak counts in respect to t=0.*

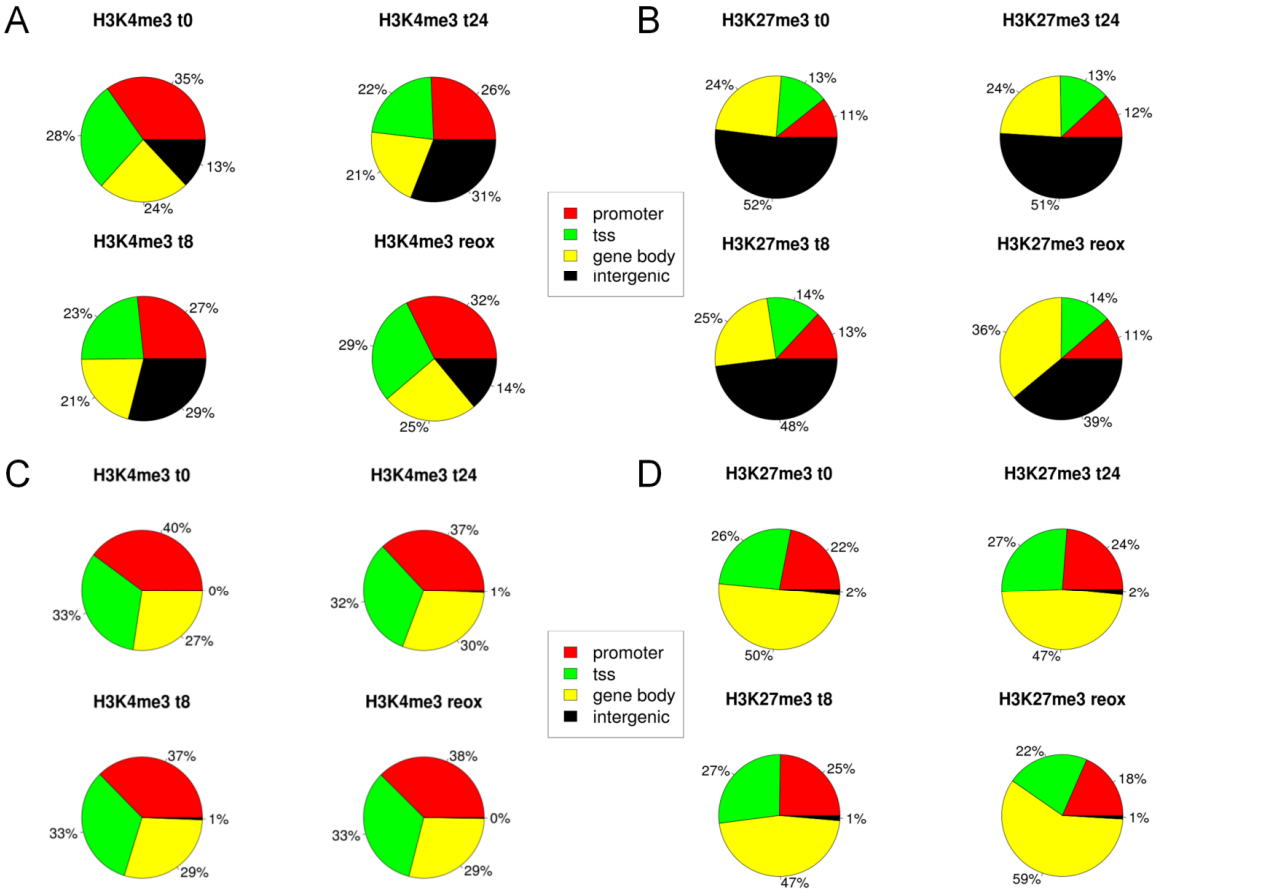


Figure S3. *Preferred gain of histone methylation at genic regions.* ***A-B****; Graphical representation of occurrence of* ***A****) H3K4me3-marks and* ***B****) H3K27me3-marks at genic (TSS, promoter, gene body) and intergenic regions.* ***C-D;*** *Graphical representation of relative distribution of sequences associated with* ***C****) H3K4me3-marks and* ***D****) H3K27me3-marks over genic (TSS, promoter, gene body) and intergenic regions.*

Fig S4

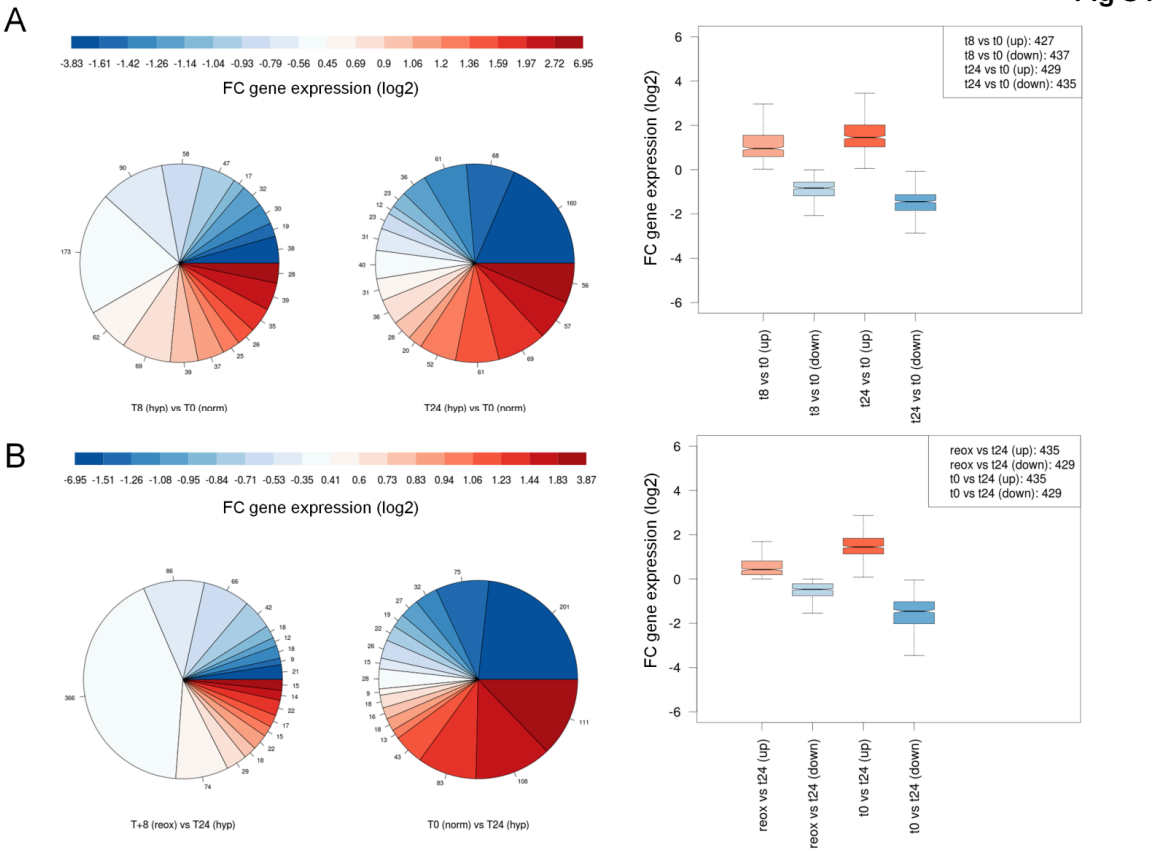


Figure S4. *Global expression changes in response to hypoxia and reoxygenation.* ***A****; Fold change (FC) of significantly regulated genes (n=864) under hypoxia (pie charts; t=8, t=24 hrs vs t=0);* ***B****; Fold change of significantly regulated genes (n=864) following reoxygenation (pie charts; t=+8 hrs (reox) and t=0 vs t=24 hrs hypoxia). Box-plots (right panels, A-B) show fold changes (FC) median expression levels for all genes up-regulated or down regulated at the indicated intervals. A regulated gene is defined as a gene with a minimum absolute expression of 100 (averaged across 3 replicates) at any experimental time point and expression level at least 2 fold changed in response to hypoxia.*

C

Hypoxia induced

GO term	# genes
apoptosis	42
cell proliferation	36
cell adhesion	25
cellular component assembly	25
regulation of cell proliferation	25
regulation of cell cycle	21
glucose metabolic process	17
response to hypoxia	16
response to peptide hormone stimulus	16
GTP catabolic process	13
anti-apoptosis	13
cellular response to hormone stimulus	13
interspecies interaction between organism	13
response to drug	12
angiogenesis	11
cell cycle arrest	11
heart development	10
response to organic cyclic compound	10
small GTPase mediated signal transduction	10
Glycolysis	9
aging	9

Hypoxia reduced

GO term	# genes
RNA processing	61
mRNA metabolic process	55
transcription from RNA polymerase II promoter	48
cell proliferation	36
regulation of apoptosis	33
ncRNA metabolic process	32
cell cycle checkpoint	30
RNA splicing, via transesterification reactions	27
S phase	27
G1/S transition of mitotic cell cycle	26
regulation of mitotic cell cycle	26
S phase of mitotic cell cycle	25
nuclear mRNA splicing, via spliceosome	25
viral reproduction	25
ribosome biogenesis	23
M/G1 transition of mitotic cell cycle	19
interspecies interaction between organisms	18
cell division	18

GO term	# genes
protein polyubiquitination	17
APC-dpt proteasomal ub-dependent protein catabolic process	16
positive regulation ubiquitin-protein ligase activity/ mitotis	16
rRNA processing	15
negative regulation ubiquitin-protein ligase activity/mitotis	15
mRNA 3'-end processing	13
antigen processing and presentation via MHC class I	12
regulation of cellular amino acid metabolic process	12
DNA strand elongation involved in DNA replication	11
DNA damage response, p53 mediated cell cycle arrest	11
methylation	11
cellular response to insulin stimulus	11
double-strand break repair	10
termination of RNA polymerase II transcription	9
mRNA export from nucleus	9
transcription elongation from RNA polymerase II promoter	8
regulation of translational initiation	8

D

Reoxygenation induced

GO term	# genes
RNA metabolic process	576
transport	403
regulation of transcription, DNA-dependent	388
cell cycle	232
protein transport	179
mRNA metabolic process	139
vesicle-mediated transport	127
chromosome organization	113
response to DNA damage stimulus	109
translation	100
positive regulation of catalytic activity	93
viral reproduction	92
cell division	89
DNA repair	74
protein ubiquitination	72
ubiquitin-dependent protein catabolic process	70
Mitosis	69
cell cycle arrest	66
ncRNA metabolic process	62
nuclear mRNA splicing, via spliceosome	51
viral transcription	50
DNA replication	45
anti-apoptosis	44
chordate embryonic development	42
G1/S transition of mitotic cell cycle	40
protein folding	40
endocrine pancreas development	40
regulation of protein localization	39
translational elongation	38
epithelium development	37
translational termination	36
regulation of translation	36
protein localization to nucleus	33
response to oxidative stress	32
G2/M transition of mitotic cell cycle	31
nucleosome assembly	31
respiratory electron transport chain	30

Reoxygenation reduced

GO term	# genes
nervous system development	293
transcription from RNA pol II promoter	229
regulation of cell death	188
regulation of transcription from RNA pol II promoter	187
regulation of apoptosis	183
neuron differentiation	172
positive regulation of transcription, DNA-dependent	163
cellular component movement	162
cell adhesion	147
intracellular protein kinase cascade	140
negative regulation of transcription, DNA-dependent	124
synaptic transmission	118
organ morphogenesis	117
cation transport	111
G-protein coupled receptor protein signaling pathway	109
positive regulation RNA pol II promoter transcription	102
blood coagulation	99
positive regulation of cell proliferation	96
central nervous system development	93
positive regulation of apoptosis	92
negative regulation of cell proliferation	89
actin cytoskeleton organization	77
negative regulation RNA pol II promoter transcription	72
Ras protein signal transduction	71
axon guidance	66
GTP catabolic process	63
regulation of cellular component movement	61
nerve growth factor receptor signaling pathway	60
negative regulation of cell cycle	59
angiogenesis	58
platelet activation	52
regulation of cell growth	51

GO term	# genes
response to hypoxia	44
positive regulation of protein phosphorylation	43
leukocyte migration	43
cellular response to insulin stimulus	42
calcium ion homeostasis	40
Rho protein signal transduction	38
sodium ion transport	35
cell junction assembly	34
protein autophosphorylation	33
energy reserve metabolic process	31
positive regulation of I-kB kinase/NF-kB cascade	28
activation of MAPK activity	27
fibroblast growth factor receptor signaling pathway	27
positive regulation of MAPKKK cascade	27
dendrite development	26
positive regulation DNA binding TF activity	26
negative regulation of protein kinase activity	25
morphogenesis of embryonic epithelium	24
neurotransmitter secretion	23
lipoprotein metabolic process	23
receptor-mediated endocytosis	21
activation of phospholipase C activity	21
regulation of angiogenesis	21
cell chemotaxis	21
cholesterol metabolic process	20
regulation of cell shape	20
skeletal muscle fiber development	20
chromatin remodeling	19
actin polymerization or depolymerization	19
antigen receptor-mediated signaling pathway	19
gluconeogenesis	18

Figure S4. *Global expression changes in response to hypoxia and reoxygenation.* ***C****; GO-analysis of genes and processes induced and down-regulated by hypoxia (t=24 hrs).* ***D****; GO-analysis of genes and processes induced and reduced by reoxygenation (t=+8 hrs).*

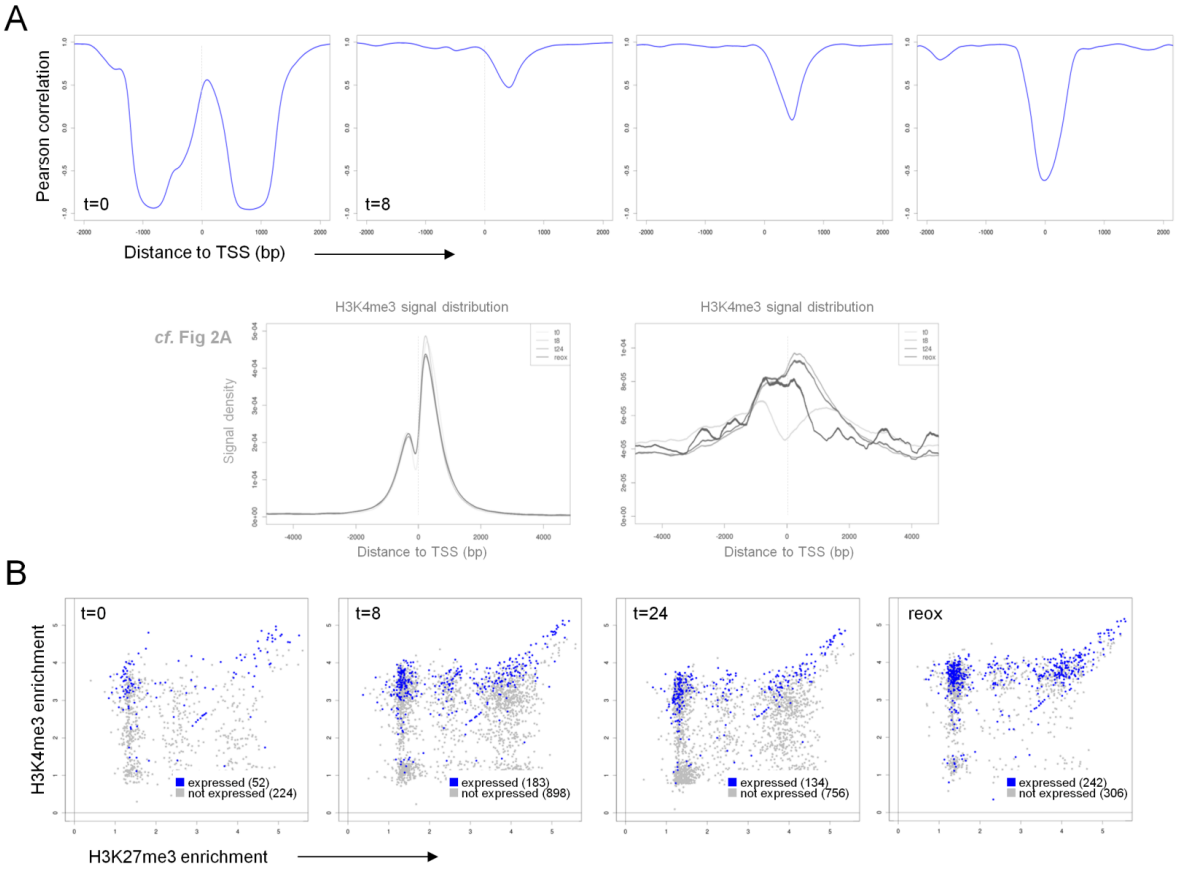


Figure S5. *Co-occurrence of H3K4me3 and H3K27me3 at TSS.* ***A****; Pearson correlation coefficients (ppc) were calculated for co-occupation by H3K4me3 and H3K27me3 in a region of +/- 2000 nt surrounding the TSS; a pcc of +1 is indicative of co-occurrence of the two marks, while a pcc of -1 indicates lack thereof. Pcc's were calculated at all indicated time points for all genes that carried associated H3K4me3 and H3K27me3-marks; lower panel - cf. Figure 2.* ***B****; double-marked genes are expressed; graphs depict genic H3K27me3-marking (x-axis) vs H3K4me3-occupation (y-axis); expressed genes are indicated with blue dots.*

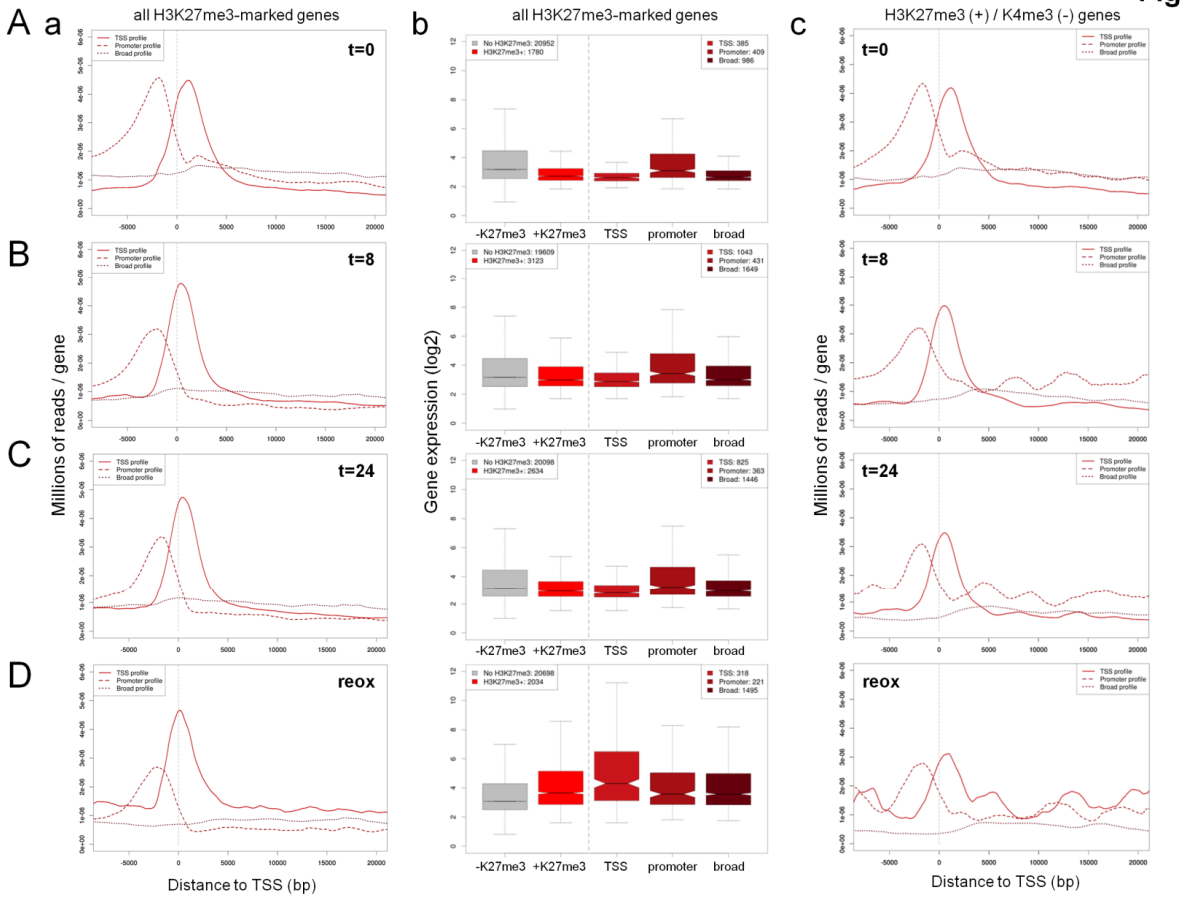


Figure S6. *Epigenetic profiles at H3K27me3-genes.* ***A-D****;* ***a*** *panels) TSS-associated enrichment profiles (average number of reads) for all H3K27me3-marked genes;* ***b*** *panels) box-plots depict median expression levels for each H3K27me3 class (a);* ***c*** *panels) TSS-associated enrichment profiles (average number of reads) for H3K27me3/single-marked genes. Comparative analysis was done at (A) t=0, (B) t=8, (C) t=24 hours hypoxia and (D) after reoxygenation (t=+8 hours).*

Fig S7

Bivalency normoxia

GO term	# genes
intracellular signal transduction	77
cell adhesion	44
axon guidance	24
negative regulation transcription by RNApol II	23
fatty acid metabolic process	17
positive regulation of cell migration	14
cell junction assembly	13
leukocyte migration	11
osteoblast differentiation	10
mononuclear cell proliferation	9
learning or memory	8
embryonic limb morphogenesis	8
BMP signaling pathway	8
heart morphogenesis	7

Bivalency 8 hours hypoxia

GO term	# genes
signal transduction	426
multicellular organismal development	419
nervous system development	214
intracellular signal transduction	206
programmed cell death	166
apoptosis	160
generation neurons	141
protein phosphorylation	126
transmembrane transport	120
cell adhesion	99
vesicle-mediated transport	97
cell migration	88
transmembrane RPTK signaling pathway	84
protein complex assembly	73
Pos regulation cell proliferation	70
central nervous system development	68
pos regulation transcription by RNA pol II	65
pos regulation of apoptosis	64
neg regulation transcription by RNA pol II	56
actin cytoskeleton organization	55
chromatin modification	54
axon guidance	51
GTPase mediated signal transduction	49
brain development	46
muscle organ development	46
response to drug	45
regulation cell migration	43
endocytosis	40
cellular amino acid metabolic process	39
muscle cell differentiation	39
angiogenesis	36
regulation cell growth	33
platelet activation	33
NGFR signaling pathway	33
response to nutrient	32
kidney development	30
exocytosis	30
morphogenesis a branching structure	28
response to hypoxia	26
osteoblast differentiation	24
intracellular receptor signaling	24
insulin receptor signaling pathway	22
carbohydrate biosynthetic process	22
bone mineralization	19
lung development	19
regulation actin cytoskeleton	19
odontogenesis	19
heart morphogenesis	18
positive regulation GTPase activity	18
activation phospholipase C activity	17
BMP signaling pathway	17
chromatin remodeling	15
embryonic limb morphogenesis	15
negative regulation of neuron apoptosis	15
regulation ion homeostasis	15
EGFR signaling pathway	14
regulation transcription RNApol II nuHR	14
electron transport chain	13
neuron migration	12
ATP biosynthetic process	12
ERK1 and ERK2 cascade	12
neg regulation canonical Wnt-R signaling	12
integrin-mediated signaling pathway	11
odontogenesis dentine-containing tooth	11
positive regulation peptidase activity	10
VEGFR signaling pathway	10

Bivalency 24 hours hypoxia

GO term	# genes
multicellular organismal development	390
signal transduction	374
regulation gene expression	224
nervous system development	197
intracellular signal transduction	172
programmed cell death	141
neuron differentiation	124
protein phosphorylation	110
response to wounding	106
cell adhesion	97
cell migration	82
central nervous system development	72
blood coagulation	64
pos regulation apoptosis	59
neg regulation apoptosis	57
pos regulation transcription by RNApol II	53
axon guidance	51
brain development	49
neg regulation transcription byRNA pol II	47
actin cytoskeleton organization	42
morphogenesis epithelium	41
regulation cell migration	41
response to drug	36
angiogenesis	34
platelet activation	34
response to metal ion	30
pos regulation of cell migration	30
cellular amino acid metabolic process	28
kidney development	27
NGFR signaling pathway	27
osteoblast differentiation	25
morphogenesis a branching structure	25
leukocyte migration	22
telencephalon development	21
regulation MAP kinase activity	21
cellular response to growth factor stimulus	21
intracellular receptor signaling pathway	20
learning or memory	19
bone mineralization	19
heart morphogenesis	18
Odontogenesis	18
BMP signaling pathway	17
bone development	16
activation phospholipase C activity	15
pos regulation of MAPKKK cascade	15
embryonic skeletal system development	15
ureteric bud development	14
embryonic limb morphogenesis	14
regulation bone mineralization	14
activation MAPK activity	13
regulation peptidyl-tyrosine phosphorylation	13
negative regulation Wnt-R signaling	13
chromatin remodeling	12
pos regulation epithelial cell proliferation	12
chondrocyte differentiation	11
regulation transcription RNApol II by nuHR	10
integrin-mediated signaling pathway	9
positive regulation bone mineralization	9
odontogenesis dentine-containing tooth	9
pos regulation osteoblast differentiation	9
VEGFR signaling pathway	9
regulation dendrite development	9
neg regulation of T cell activation	9
response to interleukin-1	9
cellular response to retinoic acid	9
mesoderm formation	8
female gonad development	8

Bivalency ESC

GO term	# genes
multicellular organismal development	93
regulation transcription, DNA-dependent	66
organ morphogenesis	49
transcription by RNApol II	46
pattern specification process	46
embryo development	44
pos regulation transcription, DNA-dpt	38
neuron differentiation	34
skeletal system development	33
central nervous system development	30
embryo development	30
pos regulation RNA pol II transcription	30
anterior/posterior pattern specification	27
neg regulation transcription, DNA-dpt	25
brain development	21
embryonic organ morphogenesis	21
neg regulation RNA pol II transcription	20
skeletal system morphogenesis	20
embryonic skeletal system development	20
cell proliferation	19
neuron development	17
embryonic skeletal system morphogenesis	15
blood vessel morphogenesis	12
dorsal/ventral pattern formation	11
neuron fate specification	11
camera-type eye development	10
regulation DNA binding TF factor activity	10
kidney development	9
limb morphogenesis	9
cartilage development	9
angiogenesis	8
forebrain development	8

Figure S7. *GO-analysis of genes and processes associated with bivalent markers at t=8, t=24 hours hypoxia; reference lists: t=0 and processes associated with known bivalent markers in embryonic stem cells.*

Supplemental Methods

Chromatin immunoprecipitation (ChIP) assays

Cells were fixed in Phosphate Buffered Saline (PBS) containing 1% formaldehyde. Cross-linking was allowed to proceed for 10 min at room temperature and stopped by a 5 minute incubation with glycine at a final concentration of 0.125 M. Fixed cells were washed twice with PBS and harvested in SDS buffer (50 mM Tris at pH 8.1, 0.5% SDS, 100 mM NaCl, 5 mM EDTA), supplemented with protease inhibitors (Aprotinin, Antipain and Leupeptin all at 5µg/ml and 1 mM PMSF). Cells were pelleted by centrifugation, and suspended in IP buffer (100 mM Tris at pH 8.6, 100 mM NaCl, 0.3% SDS, 1.7% Triton X-100, and 5 mM EDTA), containing protease inhibitors. Cells were disrupted by sonication, yielding genomic DNA fragments with a bulk size of 200-500 bp. For each immunoprecipitation, 1.2 ml of lysate was pre-cleared by adding of 35 µl of blocked protein A beads (Protein A-Sepharose/ CL-4B, GE Healthcare, Piscataway, NJ, USA; 0.5mg/ml fatty acid-free BSA, Sigma; and 0.2 mg/ml herring sperm DNA in TE), followed by centrifugation. 12 µl aliquots of pre-cleared suspension were put aside as input DNA and kept at 4°C. Samples were immunoprecipitated overnight at 4°C with primary antibodies. Immune complexes were recovered by adding 40 µl of blocked protein A beads (GE Healthcare) and incubated for 4 hours at 4°C. Beads were washed three times in 1ml of Mixed Micelle Buffer (20 mM Tris at pH 8.1, 150 mM NaCl, 5 mM EDTA, 5% w/v sucrose, 1% Triton X-100, and 0.2% SDS), twice in 1 ml of Buffer 500 (50mM HEPES at pH 7.5, 0.1% w/v Sodium Deoxycholate, 1% Triton X-100, and 1 mM EDTA), twice in 1 ml of LiCl Detergent Wash Buffer (10 mM Tris at pH 8.0, 0.5% Sodium Deoxycholate, 0.5% NP-40, 250 mM LiCl, and 1 mM EDTA), and once in 1 ml of TE. Immune complexes were eluted from beads in 250 µl elution buffer (1% SDS; and 0.1M NaHCO3) for 2 hours at 65°C with continuous shaking at 1000rpm, and after centrifugation, supernatants were collected. 250 µl elution buffer was added to input DNA samples and these were processed in parallel with eluted samples. Crosslinks were reversed overnight at 65°C, followed by a 2 hours digestion with RNAse A at 37°C and 2 hours proteinase K (0.2 µg/µl) at 55°C. DNA fragments were recovered using QIAquick PCR purification columns (Qiagen, Hilden, Germany), according to manufacturers' instructions. Samples were eluted in 75 µl EB buffer and checked for enrichment using qPCR before deep sequencing was applied on the immunoprecipitated DNA.

Protein isolation and Western blot analysis

For protein extraction cells were washed twice with cold PBS and lysed in RIPA buffer (150 mM NaCl, 1% NP-40, 0.5% w/v Sodium Deoxycholate, 0.1% SDS, 50 mM Tris at pH 8.0, 5 mM EDTA) supplemented with protease and phosphatase inhibitors (5 mM Benzamidine, 5 µg/ml Antipain, 5 µg/ml Leupeptin, 5 µg/µl Aprotinin, 1 mM Sodium Vanadate, 10 mM Sodium Fluoride, 10 mM Pyrophosphate, 10 mM ß-glycerophosphate, 0.5 mM DTT and 1mM PMSF). Lysates were subjected to two freeze-thow cycles in liquid nitrogen, followed by sonication on ice with a probe sonicator (Soniprep 150; MSE, London, UK) for 12 cycli (1 sec ON, 1 sec OFF) with amplitude 5. After 10 min centrifugation at 13.200 rpm (4°C), the supernatant was transferred to a fresh tube and protein concentration was determined using a BCA

protein assay kit (Pierce/Thermo Fisher Scientific, Rockford, IL, USA) according to the manufacturer's protocols on a Benchmark 550 Micro-plate Reader (Bio-Rad).

For immunoblotting (IB) equal amounts of protein were boiled in Laemmli buffer for 5 min and loaded on 9-15% polyacrylamide gels. Following separation by SDS-PAGE, proteins were transferred onto polyvinylidene fluoride (PVDF) membranes (GE Healthcare). Ponceau S (Sigma) staining was used to check protein transfer. Subsequently, PVDF membranes were blocked with 3.4% non-fat dry milk (Protifar; Nutricia, Zoetermeer, the Netherlands) in PBS containing 0.1% Tween-20 (pH 7.5) for 1 hour at RT, followed by an overnight incubation at 4°C with the primary antibody (see Supplementary Table 2). After extensive washing with PBS/0.2% Tween-20, membranes were probed with corresponding horseradish peroxidase conjugated secondary antibodies for 1 hour at RT: rat-anti-mouse (1:5000; DAKO, Glostrup, Denmark) and donkey-anti-rabbit (1:15.000; Jackson Lab, Bar Harbor, ME, USA), to detect monoclonal and polyclonal primary antibodies respectively. Signals were detected on autoradiograms using enhanced chemoluminescence (ECL; Pierce). Intensity of the bands was quantified with Quantity One software (Bio-Rad) and plots were generated using GraphPad Prism, version 4.03 for Windows (GraphPad Software, San Diego, CA, USA). Data was statistically analyzed by performing 2-tailed paired t-tests using Microsoft Excel. Data given is expressed as means ± standard deviation (SD) and considered significant at $p<0.05$. *, ** and *** indicate $p<0.05$, 0.01 and 0,001, respectively.

Chapter 8

General discussion

Introduction

Fast and comprehensive data processing and enhancing biological interpretation is essential in complex systems biology studies. This thesis aims to take on this challenge by constructing standardized approaches for epigenomics and cistromics data analysis. Chapter 2 evaluates the requirements of epigenomics and cistromics data analysis tools and shows how the enrichR toolkit can serve as the missing link between a variety of other available tools. Chapter 3 shows the importance of finding standardized pre-processing approaches for epigenomics and cistromics data stemming from microarray technology. Chapter 4 assesses the requirements for the processing of H3K27me3 histone modification ChIP-seq data and the power of the standardized protocol derived from this assessment. Chapter 5 proposes a work flow to integrate existing biological pathway knowledge from several sources to enable a comprehensive biological interpretation of newly generated data. And finally, chapters 6 and 7 show the power of standardized systems biology approaches when applied to cistromics and epigenomics studies. This chapter evaluates the limits of standardization and automation of epigenomics and cistromics analysis approaches and addresses the future of systems biology research in general.

Standardization of epigenomics and cistromics analysis approaches

As the number of data types increases, the number of putative analysis paths increases exponentially. It is impossible to know a priori which of these paths will be of interest. This does not only require a certain level of expertise, but also non-standard tools. When a bioinformatician creates such a tool, he is essentially developing it for other bioinformaticians, which has a large impact on the requirements: it should be flexible, modular and extendible. This means that strictly speaking, such a tool is not a standard software package, but a set of smaller tools or modules, which bioinformaticians can add to their own analysis work flows in a non-obtrusive way. In this context, a bioinformatician is anyone who can successfully incorporate these modules. Hence, there are two important requirements for standardization of analysis approaches and tools: (i) to maximize code reuse, the tools should be available as open-source; (ii) the tools should adhere to current design standards of the field, taking standard input data formats and generating standard output formats, thereby facilitating their dissemination in the scientific community.

Data from different fields all have their own analysis pipelines (**figure 1**). Standardized pipelines are for the larger part in place for transcriptomics, metabolomics, proteomics and genetics. The tools and approaches presented in this thesis in chapter 2, 3 and 4 complete the framework by adding standards for epigenomics and cistromics data analysis. Individual pipelines can be integrated into work-flow wrappers. In general, such wrappers are automated services where users upload their data and with only minimal further input, receive processed data back after some time. An example is arrayanalysis.org [1], that automates quality control, data preprocessing, statistical analysis and even biological interpretation of transcriptomics data using Pathvisio and Wikipathways [2,3]. Arrayanalysis.org makes use of open-source standards for bioinformatics analysis tools by employing R and Bioconductor packages [4].

Examples of more generalized frameworks for integrative workflows are Galaxy [5] and GenePattern [6]. Both are standardized data integration platforms that aim to make the power of bioinformatics tools accessible to scientists that lack the necessary computer programming skills. Galaxy and GenePattern put emphasis on small, easy to integrate tools that can be used as building blocks to design reproducible analysis pipelines that capture the methods, parameters and data used to produce analysis results.

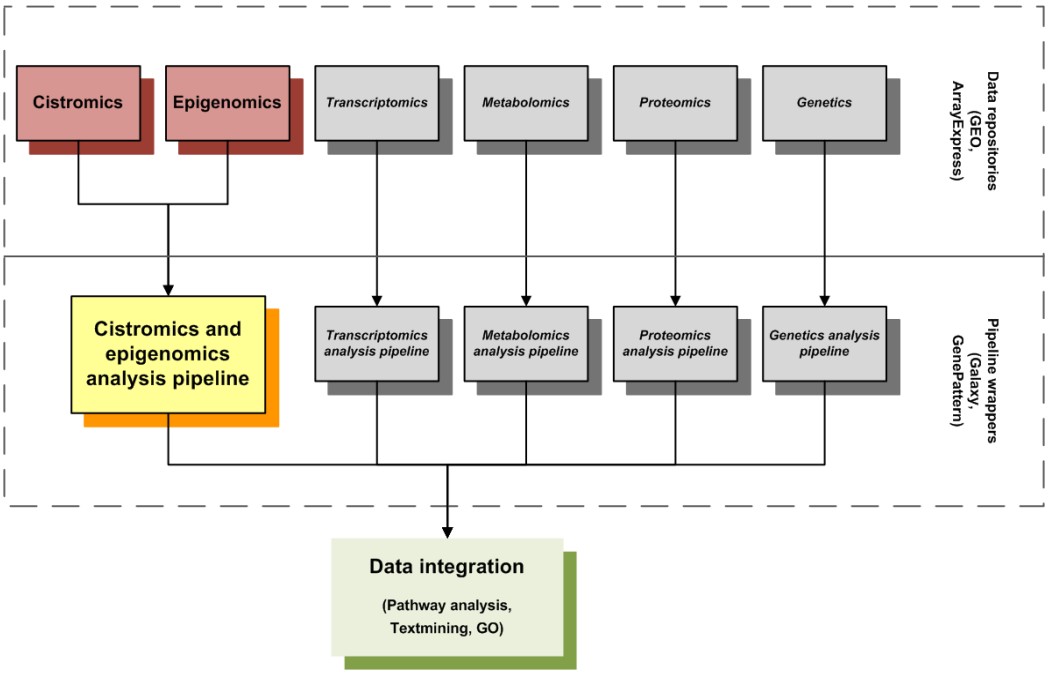


Figure 1: *Schematic of systems biology data analysis. Contributions for epigenomics and cistromics as presented in this thesis are highlighted on the left.*

One of the major threats for any bioinformatics tool is that it loses support, either from their developers or from the scientific community. There are numerous examples of online services, such as microarray quality control services RACE [7] and AMarge [8], that lost support, became inaccessible and subsequently forgotten. The enrichR toolkit presented in chapter 2 has no graphical user interface and as such may appear to target a small, specific user base with ample experience with R. However, by having as many high-level functions as possible, with suitable defaults, enrichR is able to perform complex tasks in just a few lines of code. Additionally, because enrichR is open-source and fully compatible with R and Bioconductor, it can be implemented in any framework that is compatible with this open-source standard in bioinformatics and biostatistics analysis, including Galaxy, GenePattern and arrayanalysis.org, but it can be run locally as a stand-alone tool as well. Although there are no guarantees in life nor in science, this should stimulate reusing the functionality of enrichR in future applications.

The limits of standardization of epigenomics and cistromics data preprocessing

Normalization standards are well defined for many single-channel and two-channel platforms and applications [9,10,11,12]. For two-channel microarray technology, many of these normalization approaches were originally designed for transcriptomics studies. In such studies, the two channels on the microarray comprise transcript samples, often corresponding to different conditions between which only a small amount of differential gene expression is expected. In contrast, the two channels in epigenomics and cistromics studies comprise an enriched sample, containing short DNA sequences of epigenetic modification or DNA-interacting protein locations, and a total DNA reference sample, containing short DNA sequences corresponding to the whole genome of the investigated biological system. Additionally, epigenomics and cistromics microarray data is a convolution of enriched probe and unenriched probe signals. These signals arise from sequences that are overrepresented or underrepresented, respectively, in the enriched DNA sample, corresponding to regions in the genome that are enriched and unenriched for binding of a DNA-interacting protein or presence of a specific epigenetic modification. This means that assumptions underlying transcriptomics normalization approaches will in general not hold for epigenomics and cistromics microarray data.

The results of chapter 3 suggest that using normalization approaches that do not take the unique structure of two-channel epigenomics and cistromics microarray data into account (LOWESS [12], quantile [10] and VSN normalization [13]), lower both sensitivity and specificity of identifying enriched regions in the genome compared to methods that do (Tukey's biweight scaling and T-quantile normalization [10]). Interestingly, Peng's method [14] underperforms, even though it was specifically developed for ChIP-on-chip technology and as such would appear to be the method of choice for cistromics and epigenomics microarray data. The method was originally developed for studies of male-specific lethal (MSL) complex in Drosophila [15,16] and worked well for this biological system and the microarray design for this species. These results serve as a general warning that firstly, canonical data normalization methods are very application dependent and their success depends on the technology and the biological system they were originally developed for, and secondly, that incorrect use can severely impact the reliability of the biological interpretation of the data. In conclusion, normalization methods can be standardized only for a specific technology and application, and since technology is never at a standstill, this standardization will remain a continuing effort.

Efforts to standardize pre-processing steps for high-throughput sequencing technology have recently yielded standardized work-flows and analysis tools for MeDIP-seq analysis [17] and generic MeDIP-seq, ChIP-seq and RNA-seq quality control [18]. In chapter 4, a full standardized protocol is presented for H3K27me3 ChIP-seq data analysis, tuned to studies of dynamic biological systems. There are similar approaches reported in literature for this specific histone modification, but none are fully standardized. By restricting enrichment finding to predefined regions [19,20,21,22], by not taking input sample data or an

estimated null distribution into account [20,22,23,24, 25], or by using enrichment finding approaches without settings adapted to the specific H3K27me3 enrichment profile [26,27,28], none of them offer a robust, genome-wide H3K27me3 enrichment finding approach. Additionally, the majority of the approaches skimp on data normalization, opting instead for a qualitative data analysis and comparison [24,25,26,27,28]. As such, the presented protocol is the first to enable robust, genome-wide identification of enrichment and quantitative data comparison for H3K27me3 ChIP-seq data.

Since the protocol was developed specifically to study H3K27me3 histone modifications, not all of the constructed steps are sufficiently generic to accommodate any ChIP-seq experiment. Alignment of raw reads using Novoalign, paired-end mappings and leaving out reads that align to more than one region in the genome leads to the most reliable mapping. This mapping is irrespective of the studied biological system or histone modification and hence is generic for any ChIP-seq experiment comprising paired-end reads. For example, the mappings for the H3K4me3 data presented in chapter 7 were generated using the same protocol, despite the fact that it comprises a different distribution, profile and biological function than H3K27me3 histone modifications.

The enrichment finding step in the presented protocol is completely optimized for the blanketing H3K27me3 histone modification. When the expected enrichment profile in a ChIP-seq experiment does not follow such a blanketing distribution, adapted settings will of course yield more reliable results. Similarly, data normalization depends highly on the studied biological system. The approach based on finding stable regions of enrichment in the genome is essential when studying dynamical systems, where the location and total amount of epigenetic modifications or protein binding sites is highly variable between conditions. Although this approach also works for systems with only a small amount of prospected differential sites, it is not the most efficient, especially considering that the definition of when a region is sufficiently stable are hard to define [29]. Other normalization approaches are more suited in these instances, such as scaling based on the total number of reads, or normalizing based on fitting a LOWESS model between different conditions using the number of reads ending up on a large number of predetermined genomic locations [29].

The final step before biological interpretation is summarization of the data ending up in specific regions of interest in the genome. For H3K27me3 such regions were defined previously, connecting specific regulatory characteristics to these regions in presence of this histone modification [19], but for other modifications it may be more difficult. Nonetheless, regions neighboring genic loci are a logical starting point for any ChIP-seq experiment where integration with transcriptomics data is essential.

Implications of ChIP-seq data analysis standardization for data storage

Online data analysis services are meant to automate and speed up those data processing tasks that require minimal user input, which leaves more time for the actual biological interpretation of the data. A major downside is the enormous impact such services have on the required infrastructure. In the age of high-throughput sequencing, especially data storage is a severe issue. The raw reads of a single ChIP-

seq sample can take up many gigabytes of storage, but after data pre-processing and summarization, files are smaller by a factor of at least ten [30]. Hence, it will become important in the future to consider for each stage in the work flow whether it is necessary to keep the output data to be able to answer questions arising during the biological interpretation. By standardization of high-throughput sequencing work flows, there will often be no need to repeat analysis steps once they have been completed using the state-of-the-art approach and hence many intermediary results will be redundant.

Most of the analysis steps described in the protocol for blanketing histone modifications of chapter 4 need to be performed only once during the course of a research project. Mappings obtained after aligning raw reads using Novoalign [31] are robust and can be considered final, except when a new genome build becomes available. Apart from coordinate transformations, the differences between genome builds are relatively small in genic regions, since these are well conserved. Hence, aligning to a new or different genome build will in most cases have a minimum impact on the biological interpretation of the data and unless specific intergenic regions are of interest [21,24], where insertions, deletions or crossovers may be added or removed by genome build updates, the alignment step need not be repeated.

As long as input sample reads are taken into account to robustly account for background anomalies, and the broad peak settings presented in chapter 4 are used, the enrichment finding step for blanketing histone modification ChIP-seq data also needs be performed only once. The same holds for data normalization, where for this type of ChIP-seq data, scaling based on regions with stable enrichment between conditions is a robust and deterministic approach. The only step that may have to be repeated is the data summarization step, because it depends strongly on the biological system studied. Depending on the research question, biological interpretation may be restricted to specific regions of interest, such as promoters, thereby lowering the memory and computational footprint. But inevitably, during the interpretation phase, it will become necessary to look at other regions as well.

When the above concepts are applied to the example of the H3K27me3 ChIP-seq data used for the results of chapter 4 and chapter 7, only the files containing the genomic location and intensity of enrichment peaks are essential to keep for a longer period of time. Even though the H3K27me3 ChIP-seq data presented herein is a relatively small set of four samples, this lowers the required storage capacity dramatically, illustrating that for larger datasets, removing redundant data becomes absolutely essential.

Using existing biological pathway knowledge in data analysis

Integrating existing biological pathway knowledge enables a comprehensive biological interpretation of newly generated epigenomics and cistromics data. The quality of this knowledge varies greatly between online pathway repositories. The example for the fatty acid metabolism pathways in chapter 5 shows that none of the tested repositories (BioCarta [32], Reactome [33], KEGG [34] and GenMAPP [35,36]) offer comprehensive entries, often being complementary to each other. Although Reactome offers the most complete content, it is hampered by the underlying curation process, that is concise but slow in updating

that content. Integrating the biological pathway knowledge from all these databases with current literature and expert knowledge is therefore crucial for creating canonical, comprehensive biological pathways.
The method presented in chapter 5 to improve and curate fatty acid metabolism pathways is an exhaustive but time consuming approach. Using it to create a complete map of all the processes occurring in a biological system is therefore impossible. Thankfully, there have been developments in recent years to improve pathway content on a much greater scale. Firstly, there are commercial efforts such as MetaCore from GeneGo Inc., that offer proprietary manually curated databases, that are claimed to be the most comprehensive in the field. Secondly, there are community driven approaches. Several efforts have been made to create such community driven repositories, such as the Pathway Interaction Database [37] and the possibility for submitting content to Reactome [33], but arguably the most significant one is WikiPathways [3]. WikiPathways uses the community curation principle of Wikipedia to form a comprehensive repository of canonical pathways. The plus side is that pathways are constantly curated in parallel leading to a much higher rate of improved pathways in a shorter amount of time. A downside is that the power of community curation depends largely on the size of the user base, which if not sufficiently large, will impact the size and overall quality of the content. Hence to increase support, the user-experience and the curation process itself should be as smooth and straightforward as possible. For example, future developments in this area for Wikipathways entail automated suggestions for pathway content improvement, easing the curation process by presenting information that would have otherwise taken considerable effort to salvage. Such suggestions can be retrieved from other content providers, generated by literature text-mining or visualized using omics data from public data repositories [38].

When using omics approaches, we mostly focus on presumably static images of a system under a given condition or moment in time. This is of course a simplification, as biological systems are extremely dynamic. A growing development taking place in the pathway field is the creation of mathematical models to simulate the dynamic behavior of a biological system. The Biomodels database [39] houses a plethora of such dynamic models, including models of for example signaling pathways, circadian rhythm and fatty acid transport across membranes [40]. Generally such models are created using bottom-up approaches instead of omics guided top-down approaches, measuring levels and states of a handful of proteins and metabolites of canonical processes in a controlled environment over a period of time to estimate model parameters [40]. Especially when studying cistromics, dynamic modeling can lead to new insights. In the ER-α study of chapter 5, such a model could for example comprise the ER-α signaling cascade, starting with stimulation by an estrogenic compound and ending with activated ER-α entering the nucleus to bind to target genes. This model can assist in determining the most suitable point in time to measure both ER-α binding targets using cistromics as well as gene expression using transcriptomics in a biological system stimulated by an estrogenic compound. By tuning the time of measurement, the amount of targets identified by cistromics is maximized and the amount of gene expression changes arising from secondary effects, i.e. effects unassociated with direct binding of activated ER-α, is minimized. Efforts are underway

to create an integrated dynamic model for estrogen signaling in breast cancer cells to improve the understanding of their susceptibility or resistance to endocrine therapy, but currently there are only preliminary mathematical models available of the basic decision circuits in breast cancer cells [41]. In general, dynamic modeling is beneficial for cistromics studies on any signaling cascade, such as insulin or MAPK signaling [42], that leads to the activation of a transcription factor or combination of transcription factors, and subsequent induction or repression of specific targets. As such, it is a development that will become increasingly important for cistromics and systems biology research in the coming years.

Epigenomics and cistromics to study the molecular mechanisms of cancer

There are many efforts to ease biological interpretation. The most well known is pathway analysis tools [2] to analyze data in the context of biological processes. The downside is that the pathway content used for such analyses is not always up-to-date with the latest discoveries in the field. Textmining [43] seeks to fix this problem by automated mining of the biological knowledge encoded in text documents. Regardless, analyzing results in a broad biological context will always require expert supervision. This means that the role of a bioinformatician is to develop generic tools that enable a fast and standardized analysis and that facilitate the biological interpretation by such experts. Chapters 2, 6 and 7 show applications of such standardized systems biology approaches for cistromics and epigenomics in cancer research.

Estrogenic compounds such as 17β-estradiol and tamoxifen, are among the most prescribed drugs to treat breast cancer. Both compounds activate estrogen receptor α (ER-α) which subsequently binds to the promoter of target genes, leading to up-regulation of some genes and down regulation of others at the same time. Additionally, tamoxifen can lead to different transcriptional effects in the breast compared to the endometrium. Both events may be caused by differential co-regulator recruitment. In chapter 2 and 6, systems biology approaches prove essential to dissect the molecular mechanisms underlying this differential co-regulator recruitment. Using ChIP-on-chip technology, 904 ER-α targets were identified in T47D breast cancer cells. Using gene ontology analysis and findings in literature, a sub-selection of these targets was created. For this selection of targets, it was verified that indeed up- or down-regulation of transcription correlates with the selective recruitment of co-activators or co-repressors, respectively, both for 17β-estradiol and tamoxifen stimulation, and both for breast (T47D) and endometrial carcinoma cells (ECC1).

The findings of chapter 6 are complemented by the results of chapter 2. By integrating the original ChIP-on-chip dataset with a 17β-estradiol stimulated T47D transcriptomics data from an online repository [44], we distinguished targets whose transcription is induced and targets whose transcription is repressed upon binding of ER-α. Motif analysis of the promoters of these targets showed that up-regulated targets were characterized by estrogen response element (ERE) half-sites, while down-regulated targets are characterized by full canonical ERE motifs. A gene ontology analysis on this set of targets indicated that up-regulated targets are involved in cell cycle and proliferation, in accordance with observations made

previously [45], while the down-regulated targets are mostly involved in metabolism and regulation of transcription. This suggests that the binding event of activated ER-α on the promoter is different for induced targets then it is for repressed targets. Combined, these results support the notion that recruitment of co-regulators at target gene promoters and their expression levels determine the effect of ER-α on gene expression to a large extent and additionally give a mechanism through which tamoxifen can regulate genes in opposite direction in breast and endometrial cancer cells. Applying these findings and the developed approaches on other estrogenic compounds, other cancer cell lines and clinical samples, may in the future enable predicting the action of novel estrogenic drugs in the human endometrium and other tissues and more importantly to identify patients who will benefit from hormonal therapies.

In chapter 7, ChIP-seq analysis of H3K4me3 and H3K27me3 histone modifications are combined with microarray expression data in MCF7 breast cancer cells, to study the relation between epigenetic and transcriptional changes upon hypoxic exposure and subsequent reoxygenation. Although a rapid global increase in both H3K4me3 and H3K27me3 is observed, which is largely reversed by reoxygenation, there are many subtle effects occurring on a smaller scale. The H3K4me3 enrichment profile at the transcription start sites remained largely stable under hypoxia. In sharp contrast, there is a clear gain of H3K27me3 marking around the TSS of genes already affected by the H3K4me3 modification, resulting in bivalent marking. Genes with hypoxia induced bivalency showed a significant overlap with genes known to be bivalently marked in embryonic stem cells, suggesting that hypoxia induced bivalency may result in acquisition of stem cell-like epigenetic marking. Most interestingly, this group of genes appeared resistant to the global demethylation observed during reoxygenation. Systems biology approaches enabled discovery of these subtle effects.

Epigenomics and cistromics approaches such as described in this thesis have been applied in cancer research for a long time. The key motivation is that it has long been known that cancer is characterized by system-wide deregulation [46], including disruption of epigenetic mechanisms [47,48]. Epigenetics has even become an important cornerstone of modern oncology [49]. Yet, epigenomics and cistromics approaches are being increasingly adopted in other fields as well. Although the extent of the impact is a gray area, it is well established that a pregnant woman's habits affect the health of her unborn child. When such impacts are severe, they may leave their mark on the genome of future generations through epigenetic mechanisms. Recent epigenomics studies of the Dutch Hongerwinter and the Great Chinese Famine [50] have demonstrated this, identifying small but persistent DNA methylation differences that had been passed on to later generations [51]. These small differences are genuine, but due to the nature of stable biological systems, the effect on phenotype is expected to be only slight. Hence, also in these fields, systems biology approaches are key to distinguish small methylation changes from background noise and reliably assess the impact of such changes on the condition of the system.

Future implications

It has been well established that panels of genes contribute to complex traits, with any single gene accounting for no more than a few percent of the overall variability of the trait [52]. Although classical case-control studies in the order of 1000 individuals are sufficient for identifying the genes in these panels, systems biology approaches enhance the process by providing candidate genes with a comprehensive biological context, thereby revealing key mechanisms underlying such complex traits. Yet, genomics, cistromics and epigenomics are not yet widely adopted in clinical applications. To facilitate the progression toward the clinic, standardization and clear documentation of the procedures for data preprocessing and interpretation are vital [53]. As such, the approaches presented in this thesis contribute to this development. Once such approaches have been applied in a clinical setting, marker profiles and novel drug targets can be generated for specific traits. These profiles can then be used as a screening tool to identify patients who could potentially benefit from specific treatments, or at an earlier stage to identify individuals with an increased risk at specific diseases. A recent development is people taking matters into their own hands by having themselves screened using over-the-counter tests. An example of a company offering such services is 23andMe, where any individual can submit a saliva sample to have it analyzed for over a million genetic variations for a relatively low price.

Besides offering these services, some companies additionally use the data of customers for research purposes. Although this novel way of promoting research participation is an exciting development in modern health and medicine [54], it is important to realize the impact of this development. When a subject is studied as a system, measuring metabolites and proteins on a large scale and gene expression, genetic variation and epigenetic variation genome wide instead of looking at a few individual markers and genes, one is prone to uncover results that are beyond the primary research questions but nonetheless clinically relevant for the subject. How to handle such incidental findings is not trivial. Services like 23andMe, mostly give trivial information, like eye color, hair type and genetic similarity to Neanderthals, but additionally, several indicators are given showing susceptibility to specific diseases. Although genetic variation can impact the risk of developing a serious illness, other factors also have a large influence, like nutrition, exercise and other environmental factors: nature versus nurture. It is for instance well established that diet and exercise have at least as much impact on the onset of type 2 diabetes as the presence of genetic risk factors [55,56]. Giving risk indications for serious conditions based purely on one characteristic of a system is not only ethically debatable, it would downright be incorrect and a layperson may improperly interpret such results and undertake unnecessary action [57]. The successful regulation of these developments [58] will therefore be key to the success of implementing systems biology research in everyday life.

Conclusions

Of all the steps involved in the analysis of epigenomics and cistromics data, the biological interpretation will remain the step that requires curation by human experts. The role of a bioinformatician should always

be to make that step as smooth as possible for the biologist by creating standardized, biology driven approaches. The work presented in this thesis contributes to the standardization of analysis approaches in the field of epigenomics and cistromics by providing standardized approaches for data pre-processing of ChIP-on-chip and MeDIP-on-chip microarray data, for data pre-processing of blanketing histone modification ChIP-seq data, for human expert curation of existing biological knowledge to enable a comprehensive biological interpretation, and lastly for the biological interpretation of epigenomics and cistromics data. Applying the developed approaches in biological research has yielded new insights in the mechanisms behind estrogen-dependent breast cancer and epigenetic reprogramming in hypoxic tumors. Future developments should be directed towards automating the methods and making them available as services to the scientific community. The work presented in this thesis is only a small part of the field of systems biology. All efforts in this field together have revolutionized biological research. And in this age of biology, it is only a matter of time before they claim their justly place in the clinic and everyday life as well.

References

1. Eijssen L, Jaillard M, Adriaens ME, De Groot P, Evelo CT. ArrayAnalysis.org: friendly solutions for Affymetrix microarray quality control and pre-processing. *Submitted*.
2. Van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S, Conklin BR, Evelo C: Presenting and exploring biological pathways with PathVisio. BMC Bioinformatics 2008, 9(1):399.
3. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C: WikiPathways: pathway editing for the people. PLoS biology 2008, 6(7):e184.
4. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J et al: Bioconductor: open software development for computational biology and bioinformatics. Genome biology 2004, 5(10):R80.
5. Goecks J, Nekrutenko A, Taylor J, Galaxy T: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome biology 2010, 11(8):R86.
6. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP: GenePattern 2.0. Nature genetics 2006, 38(5):500-501.
7. Psarros M, Heber S, Sick M, Thoppae G, Harshman K, Sick B: RACE: Remote Analysis Computation for gene Expression data. Nucleic acids research 2005, 33(Web Server issue):W638-643.
8. Lozano JJ, Kalko SG: AMarge: Automated Extensive Quality Assessment of Affymetrix chips. Applied bioinformatics 2006, 5(1):45-47.
9. Adriaens ME, et al. An evaluation of two-channel ChIP-on-chip and DNA methylation microarray normalization strategies. *Submitted*
10. Bolstad BM, Irizarry RA, Astrand M, Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics (Oxford, England) 2003, 19(2):185-193.

11. Shi W, Oshlack A, Smyth GK: Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. Nucleic acids research 2011, 38(22):e204.
12. Smyth GK, Speed T: Normalization of cDNA microarray data. Methods (San Diego, Calif) 2003, 31(4):265-273.
13. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M: Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics (Oxford, England) 2002, 18 Suppl 1:S96-104.
14. Peng S, Alekseyenko AA, Larschan E, Kuroda MI, Park PJ: Normalization and experimental design for ChIP-chip data. BMC bioinformatics 2007, 8:219.
15. Sural TH, Peng S, Li B, Workman JL, Park PJ, Kuroda MI: The MSL3 chromodomain directs a key targeting step for dosage compensation of the Drosophila melanogaster X chromosome. Nature structural & molecular biology 2008, 15(12):1318-1325.
16. Gelbart ME, Larschan E, Peng S, Park PJ, Kuroda MI: Drosophila MSL complex globally acetylates H4K16 on the male X chromosome for dosage compensation. Nature structural & molecular biology 2009, 16(8):825-832.
17. Huang J, Renault V, Sengenès J, Touleimat N, Michel S, Lathrop M, Tost J: MeQA: A pipeline for MeDIP-seq data quality assessment and analysis. Bioinformatics (Oxford, England) 2011.
18. Planet E, Stephan-Otto Attolini C, Reina O, Flores O, Rossell D: htSeqTools: High-Throughput Sequencing Quality Control, Processing and Visualization in R. Bioinformatics (Oxford, England) 2011.
19. Young MD, Willson TA, Wakefield MJ, Trounson E, Hilton DJ, Blewitt ME, Oshlack A, Majewski IJ: ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. Nucleic acids research 2011, 39(17):7415-7427.
20. Chopra VS, Hendrix DA, Core LJ, Tsui C, Lis JT, Levine M: The polycomb group mutant esc leads to augmented levels of paused Pol II in the Drosophila embryo. Molecular cell 2011, 42(6):837-844.
21. Rosenfeld JA, Wang Z, Schones DE, Zhao K, DeSalle R, Zhang MQ: Determination of enriched histone modifications in non-genic portions of the human genome. BMC genomics 2009, 10:143.
22. Marks H, Chow JC, Denissov S, Françoijs K-J, Brockdorff N, Heard E, Stunnenberg HG: High-resolution analysis of epigenetic changes associated with X inactivation. Genome research 2009, 19(8):1361-1373.
23. Xu H, Wei C-L, Lin F, Sung W-K: An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. Bioinformatics (Oxford, England) 2008, 24(20):2344-2349.
24. Pauler FM, Sloane MA, Huang R, Regha K, Koerner MV, Tamir I, Sommer A, Aszodi A, Jenuwein T, Barlow DP: H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. Genome research 2009, 19(2):221-233.

25. Akkers RC, van Heeringen SJ, Jacobi UG, Janssen-Megens EM, Françoijs K-J, Stunnenberg HG, Veenstra GJC: A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in Xenopus embryos. Developmental cell 2009, 17(3):425-434.
26. Kim SW, Yoon S-J, Chuong E, Oyolu C, Wills AE, Gupta R, Baker J: Chromatin and transcriptional signatures for Nodal signaling during endoderm formation in hESCs. Developmental biology 2011, 357(2):492-504.
27. Li H, Bitler BG, Vathipadiekal V, Maradeo ME, Slifker M, Creasy CL, Tummino PJ, Cairns P, Birrer MJ, Zhang R: ALDH1A1 is a novel EZH2 target gene in epithelial ovarian cancer identified by genome-wide approaches. Cancer prevention research (Philadelphia, Pa) 2011.
28. Suzuki H, Takatsuka S, Akashi H, Yamamoto E, Nojima M, Maruyama R, Kai M, Yamano H-O, Sasaki Y, Tokino T et al: Genome-wide profiling of chromatin signatures reveals epigenetic regulation of MicroRNA genes in colorectal cancer. Cancer research 2011, 71(17):5646-5658.
29. Huang W, Umbach DM, Vincent Jordan N, Abell AN, Johnson GL, Li L: Efficiently identifying genome-wide changes with next-generation sequencing data. Nucleic acids research 2011, 39(19):e130.
30. Park PJ: ChIP-seq: advantages and challenges of a maturing technology. Nature reviews Genetics 2009, 10(10):669-680.
31. Ruffalo M, LaFramboise T, Koyutürk M: Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics (Oxford, England) 2011, 27(20):2790-2796.
32. BioCarta (n.d.). Available: http://www.biocarta.com/. Accessed 6 January 2012.
33. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B et al: Reactome: a database of reactions, pathways and biological processes. Nucleic acids research 2011, 39(Database issue):D691-697.
34. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research 2011.
35. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR: GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nature genetics 2002, 31(1):19-20.
36. Salomonis N, Hanspers K, Zambon AC, Vranizan K, Lawlor SC, Dahlquist KD, Doniger SW, Stuart J, Conklin BR, Pico AR: GenMAPP 2: new features and resources for pathway analysis. BMC bioinformatics 2007, 8:217.
37. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH: PID: the Pathway Interaction Database. Nucleic acids research 2009, 37(Database issue):D674-679.
38. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, Pico AR: WikiPathways: building research communities on biological pathways. Nucleic acids research 2011.
39. Li C, Donizelli M, Rodriguez N, Dharuri H, Endler L, Chelliah V, Li L, He E, Henry A, Stefan MI et al: BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. BMC systems biology 2011, 4:92.

40. Hübner K, Sahle S, Kummer U: Applications and trends in systems biology in biochemistry. The FEBS journal 2011, 278(16):2767-2857.
41. Tyson JJ, Baumann WT, Chen C, Verdugo A, Tavassoly I, Wang Y, Weiner LM, Clarke R: Dynamic modelling of oestrogen signalling and cell fate in breast cancer cells. Nature reviews Cancer 2011, 11(7):523-532.
42. Orton RJ, Adriaens ME, Gormand A, Sturm OE, Kolch W, Gilbert DR: Computational modelling of cancerous mutations in the EGFR/ERK signalling pathway. BMC systems biology 2009, 3:100.
43. Zhou G, Zhang J, Su J, Shen D, Tan C: Recognizing names in biomedical texts: a machine learning approach. Bioinformatics (Oxford, England) 2004, 20(7):1178-1190.
44. Creighton CJ, Cordero KE, Larios JM, Miller RS, Johnson MD, Chinnaiyan AM, Lippman ME, Rae JM: Genes regulated by estrogen in breast tumor cells in vitro are similarly regulated in vivo in tumor xenografts and human breast tumors. Genome biology 2006, 7(4):R28.
45. Romano A, Adriaens M, Kuenen S, Delvoux B, Dunselman G, Evelo C, Groothuis P: Identification of novel ER-alpha target genes in breast cancer cells: gene- and cell-selective co-regulator recruitment at target promoters determines the response to 17beta-estradiol and tamoxifen. Mol Cell Endocrinol 2009, 314(1):90-100.
46. Hanahan D, Weinberg RA: The hallmarks of cancer. Cell 2000, 100(1):57-70.
47. Esteller M: Cancer epigenomics: DNA methylomes and histone-modification maps. Nature reviews Genetics 2007, 8(4):286-298.
48. Esteller M: Cancer Epigenetics for the 21st Century: What's Next? Genes & cancer 2011, 2(6):604-606.
49. Rodriguez-Paredes M, Esteller M: Cancer epigenetics reaches mainstream oncology. Nature medicine 2011, 17(3):330-339.
50. Ahmed F: Epigenetics: Tales of adversity. Nature 2010, 468(7327):S20.
51. Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES, Slagboom PE, Lumey LH: Persistent epigenetic differences associated with prenatal exposure to famine in humans. Proceedings of the National Academy of Sciences of the United States of America 2008, 105(44):17046-17049.
52. Risch NJ: Searching for genetic determinants in the new millennium. Nature 2000, 405(6788):847-856.
53. MAQC-II: analyze that! Nat Biotech 2010, 28(8):761-761.
54. Tutton R, Prainsack B: Enterprising or altruistic selves? Making up research subjects in genetics research. Sociology of health & illness 2011, 33(7):1081-1095.
55. Crandall JP, Knowler WC, Kahn SE, Marrero D, Florez JC, Bray GA, Haffner SM, Hoskin M, Nathan DM, Diabetes Prevention Program Research G: The prevention of type 2 diabetes. Nature clinical practice Endocrinology & metabolism 2008, 4(7):382-393.

56. O'Rahilly S, Barroso Is, Wareham NJ: Genetic factors in type 2 diabetes: the end of the beginning? Science (New York, NY) 2005, 307(5708):370-373.
57. Kaphingst KA, McBride CM, Wade C, Alford SH, Brody LC, Baxevanis AD: Consumers' use of web-based information and their decisions about multiplex genetic susceptibility testing. Journal of medical Internet research 2010, 12(3):e41.
58. Kaye J: The regulation of direct-to-consumer genetic tests. Human molecular genetics 2008, 17(R2):R180-183.

Samenvatting

Om biologie echt te kunnen begrijpen, moeten we het bestuderen als een systeem. Het vakgebied dat zich met dit soort onderzoek bezig houdt heet systeembiologie. Vanuit een technisch perspectief wordt een stabiel systeem gekenmerkt door feedback loops en een grote mate van redundantie. Dit dogma gaat ook op voor biologische systemen. Het aanbrengen van een kleine wijziging in het biologische systeem, zoals een veranderde activiteit van een eiwit als gevolg van een mutatie, zal over het algemeen slechts een kleine invloed hebben op het fenotype, maar zal desondanks een heel scala aan biologische pathways beïnvloeden. Dit wordt gecompliceerder bij het bestuderen van complexe aandoeningen zoals kanker, dat wordt gekenmerkt door een systeemwijde deregulering van genetische en epigenetische processen, die op hun beurt weer een heel netwerk van met elkaar verweven biologische pathways verstoren. Een huis-tuin-en-keuken voorbeeld is voeding, dat bestaat uit complexe mengsels van verschillende bioactieve stoffen, die bij inname een overvloed aan processen activeren. De enige manier om dergelijke complexe interacties te ontrafelen, is door gebruik te maken van systeembiologie benaderingen. Niet langer gericht op het meten van afzonderlijke biologische entiteiten, omvat modern biologisch onderzoek nu metingen op meerdere moleculaire niveaus met behulp van zogenaamde “omics” technologie. Dankzij deze technologie, kunnen gentranscriptie, eiwitten en metaboliet niveaus, eiwit-eiwit interacties, eiwit-DNA-interacties, genetische variatie en vele andere niveaus in één integratief biologisch kader worden geanalyseerd. De komst van deze technologie heeft er niet alleen voor gezorgd dat de hoeveelheid data in biologisch onderzoek is geëxplodeerd, maar ook dat de interpretatie van deze bergen informatie uitermate complex is geworden. Dit is waar bioinformatica om de hoek komt kijken. De grote uitdaging van de bioinformatica is computerprocedures te ontwikkelen om biologie te destilleren uit de soep van enen en nullen en is als zodanig uitgegroeid tot een van de belangrijkste hoekstenen van modern biomedisch onderzoek.

Ongetwijfeld de meest succesvolle omics technologie van de laatste tien jaar is transcriptomics, om genexpressie te bestuderen in een genoomwijde setting. Maar meer en meer beseffen we dat we niet alleen geïnteresseerd zijn in het identificeren van veranderingen in genexpressie tussen de verschillende condities, maar ook in de complexe regulatie achter deze veranderingen. In dit proefschrift staat gentranscriptie regulatie vanuit het cistroom en het epigenoom centraal. Het cistroom is gedefinieerd als de complete genoomwijde set van de cis-werkende sites op het DNA, zoals transcriptiefactor bindingsplaatsen, behorende bij een trans-werkende factor, zoals een transcriptiefactor. De complete genoomwijde set van epigenetische markeringen staat bekend als het epigenoom. Dergelijke epigenetische markeringen zijn erfelijke chromatine modificaties die los staan van veranderingen in de onderliggende DNA sequentie, die genexpressie en meer in het algemeen fenotype beïnvloeden. De meest bestudeerde epigenetische markeringen zijn DNA methylatie en chemische modificaties van histon staarten. DNA methylatie vindt meestal plaats in gebieden met een CpG di-nucleotiden. Loci met een grote hoeveelheid CpGs staan bekend als CpG eilanden, die oververtegenwoordigd zijn in die regio's in het DNA die gekoppeld zijn aan genen. Wanneer een CpG eiland wordt gemethyleerd, dan wordt door belemmering van de binding van de benodigde transcriptiefactoren en co-regulatoren de transcriptie van

het nabijgelegen gen onderdrukt. Histonen zijn eiwitten die fungeren als spoelen waar DNA omheen is gewonden, daarmee structurele eenheden vormend die nucleosomen worden genoemd. Een histon bestaat op zijn beurt uit sub-eenheden, met lange aminozuur ketens die buiten het complex uitsteken. Residuen in deze histon “staarten” kunnen chemische modificaties ondergaan die gentranscriptie kunnen induceren of belemmeren. Epigenetische markeringen hebben niet alleen hun uitwerking op de conditie van een biologisch systeem, maar dankzij hun plastische aard zijn ze ook gevoelig voor externe invloeden, zoals ziekte en het milieu. De algemene consensus is dat als dergelijke externe invloeden ernstig zijn, zoals bijvoorbeeld tijdens een periode van hongersnood, de epigenetische markeringen worden doorgegeven aan toekomstige generaties en op deze wijze bijdragen aan het aanpassingsvermogen van organismen. In deze visie worden veranderingen in de epigenetische markeringen gezien als "epigenetische littekens": belangrijke gebeurtenissen in het leven drukken hun stempel op het genoom van toekomstige generaties.

De studie van de cistroom en epigenoom in biologische systemen in een genoomwijde setting met behulp van omics technologie, wordt aangeduid als respectievelijk cistromics en epigenomics. De belangrijkste technologie om het epigenoom en het cistroom de bestuderen maakt gebruik van een combinatie van chromatine immunoprecipitatie met microarrays (ChIP-on-chip) of meer recentelijk met high-throughput sequencing (ChIP-seq). Met behulp van ChIP-on-chip of ChIP-seq technologie in combinatie met bioinformatica tools, kunnen we genoomwijde kaarten maken van verrijking voor specifieke DNA-interagerende eiwitten, histon modificaties en DNA methylatie. Dit proefschrift richt zich op belangrijke analyse vraagstukken rond cistromics en epigenomics data in elke stap van het proces, van ruwe meetgegevens tot biologische interpretatie. Het belangrijkste doel van dit proefschrift is het verbeteren van cistromics en epigenomics data analyse door de implementatie van gestandaardiseerde, integratieve systeembiologische methoden. Cistromics en epigenomics lenen zich uitstekend voor verkennende analyses, om nieuwe hypotheses te genereren die vervolgens getest kunnen worden in het lab. Voor een completere interpretatie, is integratie van verschillende omics technologieën en verschillende bioinformatica benaderingen nodig. Aangezien we bij cistromics en epigenomics geïnteresseerd zijn in regulatie van gentranscriptie, is voor een zinvolle interpretatie de integratie met transcriptomics data een eerste vereiste. Daarnaast zijn we geïnteresseerd in het ontdekken van cis-werkende motieven aanwezig in de onderliggende DNA sequenties van de geïdentificeerde verrijkte regio's, die mogelijk de binding van DNA interagerende eiwitten sturen. In hoofdstuk 2 wordt besproken aan welke eisen een cistromics en epigenomics work-flow moet voldoen en de enrichR toolkit die het mogelijk maakt om dergelijke analyses snel en grondig uit te voeren. EnrichR is gebouwd met behulp van Bioconductor, dat bestaat uit een uitgebreide, open-source collectie van bioinformatica analyse pakketten, en is zelf ook beschikbaar als open-source.
De eerste stap in elke analyse met behulp van omics technologie is het voorbewerken (pre-processing) van de ruwe metingen, om te corrigeren voor variantie van de technische aard. ChIP-on-chip en MeDIP-

on-chip is over het algemeen gebaseerd op twee-kanaals microarray technologie. Eén kanaal bevat het door immunoprecipitatie verrijkte sample, terwijl de andere een totaal DNA sample bevat, dat bestaat uit alle stukjes sequentie van het onderzochte organisme. De verhouding in intensiteit tussen de kanalen wordt gebruikt als maatregel om verrijkte regio's in het genoom te bepalen, welke overeenkomen met de DNA-interagerende eiwit bindingsites of sites met specifieke epigenetische modificaties. Deze opzet is anders dan twee-kanaals transcriptomics microarrays, waar beide kanalen transcript samples bevatten, in veel gevallen overeenkomend met twee verschillende experimentele condities. De algemene veronderstelling in transcriptomics studies is dat het merendeel van de genen qua expressie onveranderd zijn tussen condities, en dus de meerderheid van de spots op een microarray een vergelijkbaar signaal tussen de kanalen zal hebben. In epigenomics en cistromics studies gaat deze aanname echter niet op aangezien de samples in de twee kanalen zo fundamenteel verschillend zijn, hetgeen suggereert dat veel data normalisatie benaderingen ontwikkeld voor twee-kanaals transcriptomics microarray data niet geschikt zijn. Om te bepalen of dit inderdaad het geval is, hebben we de prestaties van verschillende bekende transcriptomics normalisatie strategieën geëvalueerd wanneer toegepast op ChIP-on-chip en MeDIP-on-chip data. De resultaten besproken in hoofdstuk 3 laten zien dat sommige normalisatie strategieën nog destructiever zijn dan verwacht voor de betrouwbaarheid van de biologische interpretatie.
In hoofdstuk 4 passen we wat we geleerd hebben van ChIP-on-chip en MeDIP-on-chip technologie toe om een optimaal protocol te ontwerpen voor verwerking van H3K27me3 histon modificatie ChIP-seq data. De H3K27me3 modificatie onderdrukt gentranscriptie bij aanwezigheid rond de promoter. Er zijn specifieke uitdagingen voor het bestuderen van deze histon modificatie, met name in de dynamische biologische systemen waar naar verwachting veel veranderingen in de epigenetische markeringen zullen plaatsvinden, zowel in locatie als aantal. Ten eerste zijn algoritmes voor het lokaliseren van verrijking geoptimaliseerd voor scherpe, sterk gelokaliseerde pieken, maar als gevolg van de specifieke aard van H3K27me3 modificaties om zich uit te spreiden over grotere regionen, bestaat de data ook uit meer uitgesmeerde verrijkingssignalen. Ten tweede, net als ChIP-on-chip, MeDIP-on-chip en alle andere high-throughput technologie, vereist ChIP-seq data normalisatie om te corrigeren voor technische variantie en het mogelijk te maken om de metingen van verschillende condities kwantitatief met elkaar te vergelijken. Bij het bestuderen van dynamische biologische systemen is de enige geschikte normalisatie aanpak gebaseerd op regio's in het genoom waar de verrijking stabiel is tussen de condities. Het is echter moeilijk om a priori te definiëren wanneer een regio stabiele verrijking vertoont, omdat dit sterk afhankelijk is van het bestudeerde biologische systeem. Met behulp van nieuw ontwikkelde, gestandaardiseerde procedures voor het lokaliseren van verrijking en data normalisatie hebben we deze problemen opgelost. Dit protocol is gebruikt om verrijking voor H3K27me3 histon modificaties genoomwijd in kaart te brengen voor borstkankercellen die zijn blootgesteld aan hypoxische condities.

Nadat de ruwe gegevens zijn verwerkt, komt de biologie eindelijk in beeld. Er is reeds een enorme hoeveelheid biologische kennis beschikbaar in de biologische databases in de vorm van de interacties

tussen bioactieve moleculen. Deze informatie kan worden opgeslagen als individuele biochemische interacties of als netwerken van biochemische interacties. Wanneer een dergelijk netwerk een min of meer gedefinieerd biologisch proces omvat, waarbij de afzonderlijke onderdelen aansturen op een consistent pad dat bij activering van dat proces wordt doorlopen, spreekt men van een pathway. Een belangrijke stap in interpretatie van gegevens is het integreren van kennis uit biologische databases. Een van de meest voorkomende analyse technieken die van deze kennis gebruik maakt is pathway analyse, die gen gerelateerde gegevens, zoals expressie data, visualiseren op de bestaande biologische pathways, waardoor een eenvoudige visuele interpretatie mogelijk is. Pathways zijn statische representaties van biologische processen. Echter, de kwaliteit van de pathway is sterk afhankelijk van de kennis van haar makers en de tijd die is verstreken sinds de laatste update. In hoofdstuk 5 vergelijken we de inhoud van een aantal pathways voor vetzuurmetabolisme en beoordelen hun kwaliteit door ze uit te breiden met huidige literatuur en vervolgens te laten cureren door experts. Bij deze evaluatie ontdekten we grote verschillen tussen de inhoud van databases en hebben daarnaast bijgewerkte, uitgebreide pathways van vetzuurmetabolisme gemaakt op WikiPathways.
De laatste hoofdstukken, 6 en 7, laten biologische toepassingen zien van de benaderingen besproken in de voorgaande hoofdstukken. Hoofdstuk 6 beschrijft de analyse en de resultaten van een oestrogeenreceptor α (ER-α) ChIP-on-chip dataset gericht op het uitpluizen van de moleculaire mechanismen achter de mitogene werking van oestrogeen in het menselijk endometrium en de borst, ten einde oestrogeen afhankelijke borstkanker beter te kunnen begrijpen. Hoofdstuk 7 beschrijft de resultaten van een ChIP-seq data analyse over de verrijking van activerende H3K4me3 en repressieve H3K27me3 histon modificaties. De bestudeerde systeem is een MCF7 borstkanker cellijn die is blootgesteld aan hypoxische condities, zoals een model van de effecten die optreden in de kernen van solide tumoren. Kankercellen worden gekenmerkt door een scala aan epigenetische dereguleringen. Het begrijpen van de veranderingen in de histon modificaties in zo'n dynamisch systeem is de sleutel tot het begrijpen van de moleculaire mechanismen die ten grondslag liggen aan kanker en voor de ontwikkeling van toekomstige behandelmethoden. Deze toepassingen tonen het belang van de ontwikkelde methoden en de kracht van de integratie van cistromics en epigenomics technologie in systeembiologie onderzoek.

Van alle stappen in de analyse van epigenomics en cistromics data, zal de biologische interpretatie altijd een stap blijven die supervisie door menselijke experts vereist. De rol van een bioinformaticus zou altijd moeten zijn om die stap zo soepel mogelijk te laten verlopen voor de bioloog door het creëren van gestandaardiseerde maar door de biologie gedreven aanpakken. Het werk gepresenteerd in dit proefschrift draagt bij aan de standaardisatie van analyse benaderingen op het gebied van epigenomics en cistromics door middel van gestandaardiseerde methoden voor data pre-processing van ChIP-on-chip en MeDIP-on-chip microarray data, data pre-processing van H3K27me3 histon modificatie ChIP-seq data, voor het integreren en verbeteren van bestaande biologische kennis, en ten slotte voor de interpretatie van epigenomics en cistromics gegevens in een biologische context. Toepassing van de

ontwikkelde aanpak in biologisch onderzoek heeft geleid tot nieuwe inzichten in de mechanismen achter de oestrogeen afhankelijke borstkanker en epigenetische herprogrammering in hypoxische tumoren. Het werk gepresenteerd in dit proefschrift is slechts een klein deel van het vakgebied der systeembiologie. Alle inspanningen binnen dit vakgebied hebben een revolutie teweeggebracht in biomedisch onderzoek en het is slechts een kwestie van tijd voordat deze ontwikkelingen hun weg zullen vinden naar de kliniek en ons dagelijks leven.

Acknowledgements

Despite appearances, a PhD is not something you do on your own. I would like to start by thanking my promoter, Prof. Frederik-Jan van Schooten, my co-promoter Dr. Chris Evelo and the members of the thesis committee, Prof. Frans Ramaekers, Prof. John Mathers, Prof. Michael Müller, Prof. Harald Schmidt and Dr. Danyel Jennen for reading and judging the work presented in this thesis.

It doesn't feel like four years have passed since I first set foot in the BiGCaT group. That group became the official Bioinformatics Department of Maastricht University not that long ago and in the process saw many changes. Some good, some less so. To all the BiGCaTs, past and present: thanks for all the laughs, the fun, the beers and of course the science. Chris, jij was voor mij een ideale baas. Je hebt me altijd de vrijheid gegeven om mijn eigen ding te doen, en er telkens voor gezorgd dat de omstandigheden optimaal waren om een stap verder te komen. Doordat je me meteen vanaf het begin overal mee naartoe sleepte, had ik al vroeg in mijn PhD toegang tot een aantal grote netwerken. Daar heb ik niet alleen heel plezierige contacten aan over gehouden maar ook wetenschappelijk veel profijt van gehad. Lars, als er zoiets is als een wetenschappelijke "Wingman", dan ben jij 't. Als werkpaard van BiGCaT heb je het niet altijd makkelijk en meestal druk, maar had desondanks altijd een luisterend oor. Jij hebt me uiteindelijk in contact gebracht met de mensen van Moleculaire Genetica, met een aantal zeer fijne samenwerkingsprojecten tot gevolg. Ik zie ons meer als vrienden dan als collega's, en dat is iets speciaals. Zie je snel op een zonnig terras in het zuiden of in een gezellig cafeetje aan de gracht! Magali, I remember that we were so satisfied with our newly invented "REVU plot", that we hung it up on the wall opposite our desks, so we could stare at it each time we needed to relax. Although the normalization paper eventually ended up giving everyone involved grey hairs, in the end we finally succeeded. Hope to see you soon! Jahn, thanks for the ukulele lessons, although you know deep inside that I only used it to impress Wieteke. Your wedding and the trip to your new place in Berlin were fantastic. Susan, Andra, Thomas, bedankt voor de gezelligheid op momenten dat onze wetenschappelijke en niet-wetenschappelijke paden kruisten. Martijn, bedankt voor de game tips en de vele gezellige biertjes door de jaren heen. Tina, Tini, and George, thanks for the fun times fixing ice-cream machines around the world (and who needs Australia anyway). Jos, bedankt voor de goede zorgen door de jaren heen (zoals bier op kosten van de zaak), het (niet altijd gevraagde) carrière advies en het regelen van al mijn papierwerk. Arie en Stan, jullie hebben me in het prille begin bij BiGCaT in aanraking gebracht met R en jullie aanstekelijke gevoel voor humor ("To R or not to be"). Adem, you were probably more iPhone crazy than anyone I know, although I've heard rumors of you switching to the competitor. I really enjoyed our talks over coffee and it was a pleasure to be at your wedding. Wish you all the best with the remainder of your PhD and your scientific career. Charly, I promised you so many beers over the years, that you can start a bar with me a as a supplier. Unfortunately, I was not planning a career in the liquor business, so instead, I propose you pass by any time you feel like having one of those many, many beers. And finally

to all current and future PhD students of BiGCaT: yes, all the rumors are true. Lars will do pretty much any work in exchange for beer.
Wie de studenten heeft, heeft de toekomst? Laura, Sarah, Stefan, Bo, Maarten, Ferry, Jonas en Pieter, bedankt voor jullie inzet en vooral de gezelligheid. Laura, bedankt voor de goede zorgen en de "bezorgservice". Wanneer kom ik weer eten? Sarah, bedankt voor de vele koekjes. Maarten, bedankt voor de vele biertjes. Bo, if you decide to do a PhD, be sure to let me know.

Maar de BiGCaTs zijn zeker niet de enige mensen van Universiteit Maastricht die ik moet bedanken. De laatste twee jaar heb ik een bijzonder fijne samenwerking gehad met de mensen van Moleculaire Genetica, specifiek Peggy, Willem, Frank en Guus. Peggy, je was de afgelopen twee jaar wetenschappelijk steun en toeverlaat en voor een kortstondige periode in maart 2012 zelfs even mijn persoonlijke secretaresse. Ik heb veel plezier beleefd aan onze "Seqs and the City" meetings, en wetenschappelijk gezien rolde er aan het eind ook nog eens mooie dingen uit. We promoveren nu bijna op hetzelfde moment en dat is toch wel bijzonder. Willem, jij bleek al snel de ideale sparring partner te zijn bij onze meetings. Jouw gedrevenheid weet me altijd te motiveren, zelfs op momenten dat ik door de bomen het bos niet meer zie. Daarnaast was het fijn om te zien dat naarmate de tijd vorderde, je "the Peggy-Michiel axis" het werk steeds meer zelfstandig toevertrouwde. Ik hoop in de toekomst nog vaker met je samen te werken. Everyone from the Toronto labs (Brad, Marianne, Michelle, Tim), thank you for the help with the ChIP-seq work and the pleasant conference calls. Brad, thank you for making this collaboration possible, for your kind words and your advice.
Andrea, thank you for your kind collaboration through the years. You really helped me a lot during my Master thesis and my PhD. I believe we still owe each other a few beers, so let's plan a meeting soon!
Nard, je bleek vaak de stem der rede in een kakofonie van e-mails en commentaren. Je bent een van de meest getalenteerde wetenschappers die ik ken, en nog best gezellig ook (knipoog). Hoop in de toekomst nog vaak met je samen te werken!
Is a scientist only as good as the whisky he drinks? Perhaps... Anyway, Lars, Mike and Jahn, thanks for the great times during our after (!) work whiskies. The fact that I am yet to find replacements says a lot. It's also problematic, because whisky is meant for sharing, so my bottles are piling up. Maybe we should start planning that trip to Scotland after all.
Maar een universiteit is natuurlijk nergens zonder goede faciliteiten voorziening. Hier vallen natuurlijk ook de heren van ICTS onder. Aan jullie de dank voor het blokkeren van mijn ethernet verbinding, ten behoeve van het inperken van mijn van school uit meegeven en door Bram Cohen beroemd gemaakte vorm van procrastinatie. Mijn trage WiFi connectie bleef gelukkig gespaard, zodat ik mij in de laatste fase van mijn PhD slechts volledig kon richten op mijn werk.

I became a NuGO member when I was still a Master's student, but it proved to be an important membership. I would like to thank the Nutritional Epigenomics Focusteam and specifically the wonderful

people from Newcastle University (Jill, John, Sofia, Caroline) and IFR (Nigel, Lawrence) who I ended up collaborating with. John, thanks for giving me the opportunity to be part of the excellent focus team and making the visits to Newcastle possible. It was an absolute pleasure to work with you and an honor to have you in my PhD committee. Nigel and Lawrence, you were the first to give me DNA methylation microarray data to, in Nigel's words, "play around with". Although it proved to be a pain to handle, I still consider my experience with this data invaluable. Thank you for that. Lawrence, man, thanks for the fun times, and let me know if you need anymore cheese. Jill, we have worked together for a long time now, starting when I was still a student and got "lost" during my first NuGO Nutritional Epigenomics Focus Team meeting at Stansted Airport. Although some would say that we spend most of our time drinking and gambling, I know better. I hope to work with you again in the future. And maybe gamble and drink, but only if you insist.

Thankfully, it wasn't all serious business in NuGO. Who could forget The Notorious NuGO Berlin 2008 Crew! Marijana, Katharina, Lars ("the driver"), Lars ("absolutely not the driver"), Michel, Otto and Lawrence, our improvised trip with an apparently "borrowed" TNO van to the lively centre of Berlin made NuGO week 2008 the best one ever, although I have to admit that I've never hugged a lantern since. Claus, thanks for your invaluable advice over the years (especially for the normalization paper) and your kind spirit. Tony, thank you for the NBXs and whiskies.

Otto! I think I have never laughed as much in my entire life as during the period when you came to visit BiGCaT. Although some might say that we spend most of our time smoking and drinking, others might say that I know better. What I really know, is beyond reconstruction. Although I must admit that I've never seen a monkey on a computer, a vague memory says otherwise, although details have escaped me, or confused themselves with dreams (my sincere apologies for the John Mayer quote). Yet, in my increasingly rarer moments of clarity, some of these details simmer through, about a dog who wrote a Blog, who lived in an endless garden with a mysterious house at what, to the layman's eye, would seem like the end of said garden, which is of course impossible, for the garden is endless. It's an endless garden. Neverending. And I think everyone has at one point or another heard the whispers in the hallways about 4-dimensional Friday figuroids in a hyper-real version of purple-haze Excel. Go figure, or somezing like zhat. Thankfully Terhi is always there to catch you when you fall. Terhi, you are one of the kindest persons in the world, with the patience of an angel. During my satisfactory (= Not-Sad-Is-Factory, or 'Appy-Is-Factory) stay in Kuopio, I found out that Finland is a very warm place, even in winter times. I hope to visit you both many times in the future and hope you will do the same.

Bas, Philip, Guido. Hoewel sommige mensen beweren dat wij elkaar alleen maar beledigen, weet ik wel beter. Bas en Philip, onze etentjes door de jaren heen, hielden me op de been (deze zin rijmt met opzet). Hoewel onze eerste gezamenlijke business venture afgezien van enkele taalkundige hoogstandjes weinig heeft opgeleverd, kunnen we in ieder geval zeggen dat we hebben gelachen. Bas, bedankt voor de steun tijdens de laatste loodjes, en in de zon alle heerlijke broodjes (wederom, met opzet). Guido, jij bent al jaren, officieus sinds mijn 4e, m'n trouwe steun en toeverlaat en dat zal hopelijk ook altijd zo blijven. Op

dit moment ondervind je zelf de ups en downs van een promotieonderzoek, en zult over een paar jaar op hetzelfde punt zijn aanbeland als ik nu. Gouden tip: denk op moeilijke momenten aan het pluchen aapje, dat relativeert werkelijk alles. Philip, bedankt voor je zijn, door de jaren heen. Het kostte af en toe wat moeite om je op de been te krijgen casu quo te houden, maar uiteindelijk ben je er toch gekomen. Wanneer gaan we een biertje doen?
Waar is een man zonder muziek, zonder een band? Erik, Tim, Steef, Roy, bedankt voor alle muziek door de jaren heen en alle daarmee gepaard gaande inspirerende biertjes en whisky's. Erik en Tim, bedankt voor de hulp tijdens al mijn verhuizingen door de jaren heen. Erik, buurman, ik denk met plezier aan onze tijd in Eindhoven terug, inclusief de korte terugkeer in de vorm van een vakantie in de zomer van 2010 (Luske!!!?!?!!!). Hoewel je in België woont, kom ik graag een keer jammen in je grote kelder.
Lambèr en Angèle, bedankt voor alle goede zorgen bij de eindexamentrainingen. De mogelijkheid om ervaring op te doen als docent was erg spannend, maar vooral ook erg leuk. Familie Eggen, bedankt voor de steun in het prille begin. Soms lopen zaken nu eenmaal niet zoals je verwacht, maar ik heb er desondanks goede herinneringen aan overgehouden. Andreas en Jenny, bedankt voor alle kunstzinnige en muzikale inspiratie. Ik hoop dat het ondanks onze drukke agenda's dit jaar eindelijk gaat lukken om weer eens af te spreken.

Pap, mam, broers, nichten, neef, ooms, tantes, opa en oma's, een promotie traject is zwaar, ook voor een zondagskind. Ik heb de afgelopen vier jaar menig slapeloze nacht beleefd, en dat kan ik zelfs als langslaper nooit meer bijslapen. Het heeft me als persoon sterker gemaakt, maar zeker ook veranderd. Pap en mam, bedankt voor alle steun en voor de opvang in een periode waarin het even allemaal iets minder ging. Max en Stephan, betere broers kan deze broer zich niet voorstellen. De vakantie naar Japan en onze gezamenlijke vakantie met Erik in Eindhoven (Luske!!!?!?!!!) waren bijzonder relaxt. Oma, bedankt voor alle zoetig- en zoutigheden en de "oma peptalks". Ik beloof dat ik iedere maand zal blijven bellen.
Pluk, njipnjipnjipnjip, njipnjip, njip? Njip! Wieteke, bedankt voor je geweldige steun en liefde. Sommige mensen zullen beweren dat wij als middelsten en zondagskinderen het maar makkelijk hebben met zijn tweetjes. En daar hebben ze volkomen gelijk in, ondanks mijn cafetière fetish en jouw curieuze obsessie met dode dieren. Zondagskinderen, unite! Hoewel het afronden van dit werk vaak als een einde gezien wordt, voelt het voor mij meer als het begin van het echte avontuur. En ik heb het sterke gevoel dat dat helemaal goed gaat komen.

Curriculum vitae

Michiel Emanuel Adriaens was born on 18 December 1983 in the sunny south of the Netherlands, in an old miner's town by the name of Heerlen. From an early age on, when asked about what he wanted to become later in life, he always replied inventor ("uitvinder") or cook ("koker"), although this was mostly fuelled by his urge to recreate Willie Wortel's Lampje, and have a nice meal while doing so. After finishing the Gymnasium at Bernardinus College in Heerlen in 2002, Michiel went on to study Biomedical Engineering at the Technische Universiteit in Eindhoven. Initially determined to become a biochemist, during practical sessions in the lab he found out that despite his cooking ambitions, he was not the type for following exact recipes. Bioinformatics and mathematical modeling had become more his thing, and after an internship at the Bioinformatics Research Centre at the University of Glasgow in 2006, he knew that he wanted to become a researcher. During his final Master's project, he became part of the BiGCaT Bioinformatics team at Maastricht University, working on analysis approaches for DNA methylation microarrays. This shifted to the more general epigenomics and cistromics analyses during his PhD, which he started at the same department in 2008. Currently, Michiel is experiencing the difference in climate at the Academic Medical Center in Amsterdam, hunting for genes underlying susceptibility for acute cardiac death. And despite his career choices some might say that additionally, he has become quite the inventive cook.

List of publications

1. Prickaerts, P, M.E. Adriaens, T. van den Beucken, V. Dahlmans, M. Chan-Seng-Yue, B. Wouters* and J.W. Voncken* (2012). "**Epigenetic signatures in the context of gene regulation under hypoxia.**" *[manuscript in preparation]*
2. Adriaens, M.E., W. Arindrarto, A. Romano, L.M.T. Eijssen, C.T.A. Evelo (2012). "**Systems biology approaches for ChIP-on-chip and DNA methylation microarray data.**" *[manuscript in preparation]*
3. Eijssen, L.M.T., M. Jaillard, M.E. Adriaens, P. de Groot and C.T.A. Evelo (2012). "**ArrayAnalysis.org: automated friendly solutions for Affymetrix microarray quality control and pre-processing.**" *[manuscript in preparation]*
4. Adriaens, M.E., P. Prickaerts, M. Chan-Seng-Yue, T. Beck, J.W. Voncken, B. Wouters and C.T.A. Evelo (2012). **"Capturing ChIP-seq profiles of H3K27me3 in dynamic biological systems"**. *[submitted]*
5. Mykkänen, O.T., G. Kalesnykas, M.E. Adriaens, C.T.A. Evelo, R. Törrönen and K. Kaarniranta (2012). **"Bilberries potentially alleviate stress related retinal gene expression induced by a high-fat diet in mice"**. *[submitted to Molecular Vision]*
6. Kolehmainen, M., O. Mykkänen, P. Kirjavainen, T. Leppänen, E. Moilanen, K. Hanhineva, J. Paananen, M.E. Adriaens, D. Laaksonen, M. Hallikainen, R. Puupponen-Pimiä, L. Pulkkinen, H. Mykkänen, H. Gylling, K. Poutanen and R. Törrönen (2012). "**Bilberry products modify low grade inflammation in study participants with features of metabolic syndrome.**" *[submitted to American Journal of Clinical Nutrition]*
7. Kubben, N., M.E. Adriaens, W. Meuleman, W. Voncken, Y. Pinto, B. Van Steensel and T. Misteli (2012). "**Genome-wide disruption of chromatin-lamina interactions by progerin.**" Chromosoma.
8. Adriaens, M.E., M. Jaillard-Dancette, L. Eijssen, C. Mayer and C.T.A. Evelo (2012). "**Normalization strategies for two-channel ChIP-on-chip and DNA methylation microarray technologies.**" BMC Genomics.
9. Romano, A., M.E. Adriaens, S. Kuenen, B. Delvoux, G. Dunselman, C.T.A. Evelo and P. Groothuis (2010). "**Identification of novel ER-alpha target genes in breast cancer cells: gene- and cell-selective co-regulator recruitment at target promoters determines the response to 17beta-estradiol and tamoxifen.**" Mol Cell Endocrinol 314(1): 90-100.
10. Orton, R.J., M.E. Adriaens, A. Gormand, O.E. Sturm, W. Kolch and D.R. Gilbert (2009). "**Computational modelling of cancerous mutations in the EGFR/ERK signalling pathway.**" BMC Syst Biol 3: 100.
11. McKay, J.A., M.E. Adriaens, D. Ford, C.L. Relton, C.T.A. Evelo and J.C. Mathers (2008). "**Bioinformatic interrogation of expression array data to identify nutritionally regulated genes potentially modulated by DNA methylation.**" Genes Nutr 3(3-4): 167-71.
12. Adriaens, M.E., M. Jaillard, A. Waagmeester, S.L.M. Coort, A.R. Pico, C.T.A. Evelo (2008). "**The public road to high-quality curated biological pathways**." Drug Discov Today 13(19-20): 856-62.